
**Universidade Estadual de Campinas**

Faculdade de Engenharia Mecânica

**André Luiz Pontara Torres**

# Prediction of pre-salt electrofacies from well log data in Santos basin- Brazil, by models based on Artificial Intelligence algorithms

## Predição de eletrofácies do pré-sal a partir de dados de perfis de poços da bacia de Santos - Brasil, por modelos baseados em algoritmos de Inteligência Artificial

**CAMPINAS**

**2022**

**André Luiz Pontara Torres**

# Prediction of pre-salt electrofacies from well log data in Santos basin- Brazil, by models based on Artificial Intelligence algorithms

## Predição de eletrofácies do pré-sal a partir de dados de perfis de poços da bacia de Santos - Brasil, por modelos baseados em algoritmos de Inteligência Artificial

Dissertação apresentada a Faculdade de Engenharia Mecânica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência e Engenharia de Petróleo

Dissertation presented to the Faculty of Mechanical Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Science and Petroleum Engineering

**Supervisor/Orientador: Alessandro Batezelli**

Este exemplar corresponde à versão da dissertação apresentada na defesa da dissertação de mestrado pelo aluno André Luiz Pontara Torres e orientada pelo Prof. Dr Alessandro Batezelli

**Campinas**

**2022**

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

T363p

Torres, Andre Luiz Pontara, 1990-
Prediction of pre-salt electrofacies from well log data in Santos basin - Brazil, by models based on Artificial Intelligence algorithms / Andre Luiz Pontara Torres. – Campinas, SP : [s.n.], 2022.

Orientador: Alessandro Batezelli.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Mecânica.

1. Inteligência Artificial. 2. Eletrofácies. 3. Carbonatos. 4. Pré-sal. I. Batezelli, Alessandro, 1972-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Mecânica. III. Título.

Informações para Biblioteca Digital

**Título em outro idioma:** Predição de eletrofácies do pré-sal a partir de dados de perfis de poços da bacia de Santos - Brasil, por modelos baseados em algoritmos de Inteligência Artificial.
**Palavras-chave em inglês:**
Artificial Intelligence
Electrofacies
Carbonates
Pre-salt
**Área de concentração:** Reservatórios e Gestão
**Titulação:** Mestre em Ciências e Engenharia de Petróleo
**Banca examinadora:**
Alessandro Batezelli [Orientador]
Gelvan André Hartmann
Leidy Alexandra Delgado Blanco
**Data de defesa:** 21-02-2022
**Programa de Pós-Graduação:** Ciências e Engenharia de Petróleo

**Identificação e informações acadêmicas do(a) aluno(a)**
- ORCID do autor: https://orcid.org/0000-0002-2740-9495
- Currículo Lattes do autor: http://lattes.cnpq.br/6918479355937620

**UNIVERSIDADE ESTADUAL DE CAMPINAS**
**FACULDADE DE ENGENHARIA MECÂNICA**

**DISSERTAÇÃO DE MESTRADO ACADÊMICO**

# Prediction of pre-salt electrofacies from well log data in Santos basin- Brazil, by models based on Artificial Intelligence algorithms

# Predição de eletrofácies do pré-sal a partir de dados de perfis de poços da bacia de Santos - Brasil, por modelos baseados em algoritmos de Inteligência Artificial

Autor: André Luiz Pontara Torres

Orientador: Alessandro Batezelli

A Banca Examinadora composta pelos membros abaixo aprovou esta Dissertação:

**Prof. Dr. Alessandro Batezelli**
**DGRN/Instituto de Geociências/Unicamp**

**Prof. Dr. Gelvan André Hartmann**
**DGRN/Instituto de Geociências/Unicamp**

**Prof. Dr. Leidy Alexandra Delgado Blanco**
**Instituto Colombiano del Petróleo Juan José Turbay, ICP-Ecopetrol**

A Ata de Defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 21 de fevereiro de 2022.

*Este trabalho é dedicado aos meus pais, aos meus irmãos*
*e a todos os familiares e amigos que sempre estiveram ao meu lado.*
*This work is dedicated to my parents, my siblings,*
*and all the relatives and friends who have always been by my side.*

# Acknowledgements

This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Process 301200/2017-3.

I would like to thank the University of Campinas, for being this immense place of knowledge that allows us to acquire new knowledge at each meal or walk.

I wish to thank the Brazilian National Agency of Petroleum for providing the data set that was used in this work.

I also to thank the Faculty of Mechanical Engineering(FEM) and the Petroleum Engineering Department (DEP) for contributing with all physical structure, buildings, and laboratories, which enabled the project to be carried out.

I also to thank the Institute of Geoscience(IG) and the Petroleum Engineering Department (DEP) for contributing with all physical structure, buildings, and laboratories, which enabled the project to be carried out.

I would like to thanks my supervisor, Alessandro Batezelli, for all the advice, suggestions, text corrections, meetings, patience, and confidence.

I am grateful to my firsts advisors, Emilson and Beto for teaching and support me at the beginning of the academic life with the scientific initiation projects that were entrusted to me.

I am immensely grateful to my parents and siblings for the support during my whole life. It is immeasurable and can not be described in words how grateful I am to have them in my life.

I also to thank my friends, Dan, Xande, Betão, Marcel, Breno, Teteus, Fabinho, Egs, Gui, and Hariel for the good and bad moments that have been present on my side, together with my family, you are my "backup", thanks for walking with me.

In conclusion, I would like to thank each person who passed through my life and left a little piece of their world to compose my understanding of life.

*It is only with the heart that*
*one can see rightly;*
*what is essential is invisible*
*to the eye.*
*(The Little Prince)*

# Resumo

Classificar e predizer conjuntos de dados tem sido um dos maiores desafios dos tempos modernos, principalmente porque a quantidade de dados adquiridos tem aumentado significativamente nos últimos anos. Na geologia do petróleo, o processo de analisar um grande volume de dados de perfis de poços com o objetivo de extrair propriedades do reservatório por abordagens manuais é uma tarefa bastante difícil e custosa. Assim, métodos e algoritmos que se propõem a classificar e predizer esses dados tem se tornado grandes aliados para compreender melhor a grande quantidade de dados e informações coletadas. Entre os algoritmos usados para classificar e predizer grandes bancos de dados, algoritmos de aprendizado supervisionado e não-supervisionado tem ganhado destaque nos últimos anos. Algoritmos de aprendizagem supervisionada são algoritmos que inicialmente precisam ser ensinados por um conjunto de dados previamente rotulados, enquanto que algoritmos de aprendizagem não supervisionada não necessitam que os dados sejam previamente rotulados sendo esse tipo de algoritmo capaz de trabalhar com dados não rotulados. O Support Vector Machine (SVM), Redes Neurais e K-means são exemplos de algoritmos de aprendizagem supervisionada e não-supervisionada que foram usados na pesquisa atual. Eles são algoritmos que através de cálculos matemáticos robustos são responsáveis em auxiliar na classificação e predição dos banco de dados. No trabalho foi utilizado os algoritmos Support Vector Machine (SVM), Redes Neurais e K-means para classificar e predizer diferentes eletrofacies presentes no banco de dados da pesquisa. O banco de dados da pesquisa foi construído com os dados de perfil de poço fornecidos pela Agência Nacional de Petróleo (ANP). Os perfis geofísicos que foram selecionados para alimentar nossos modelos como dados de entrada foram, Raios gama, Neutrão, Resistividade, Saturação em água, Densidade, Porosidade, Velocidade de onda P e S. Entre os algoritmos de aprendizagem supervisionada utilizados na pesquisa (SVM e Redes Neurais) fica claro de acordo com a etapa de validação estatística dos modelos (matriz de

confusão e validação cruzada K-dobras) que o modelo de Redes Neurais apresenta um melhor desempenho em todos os cenários quando comparado ao Support Vector Machine. A acurácia final do modelo baseado no SVM avaliada pelo métodos matriz de confusão e validação cruzada K-dobras foram respectivamente 92.9% e 86.6%. As mesmas medidas para o modelo baseado em Redes Neurais alcançou resultados de 96.6% e 96.9% respectivamente. A performance melhor pode ser verificada por meio de outros parâmetros, tais como, precisão, recall e pontuação F1. O único ponto no qual o Support Vector Machine têm significativas vantagens em relação as Redes Neurais é no tempo de execução. Em quanto que o modelo baseado no SVM consegue predizer as eletrofacies em 0.096s, o modelo baseado em Redes Neurais leva 41.45s para concluir suas predições. Ou seja, o modelo baseado no SVM mostra-se cerca de 432 vezes mais rápido do que o modelo baseado em Redes Neurais. Os perfis de poço preditos tiveram duas eletrofácies preditas nos algoritmos de aprendizagem supervisionada, mostrando uma convergência entre os modelos. Os modelos aplicados na pesquisa com algoritmos de IA mostraram-se robustos para classificar e predizer eletrofácies, sendo promissores no processo de automatização do processo de classificação de eletrofácies. A pesquisa mostra novas maneiras de classificar e predizer eletrofácies e tenta abrir novas possibilidades para algoritmos de Inteligência Artificial no mundo da Geociências.

**Palavras-chave**: Inteligência Artificial; Eletrofacies; Carbonatos; Pré-sal Brasileiro

# Abstract

Classifying and predicting data sets has been one of the greatest challenges in modern times, mainly because the amount of data acquired has increased a lot in the last years. In petroleum geology, analyzing a large volume of well log data to extract reservoir properties by manual approaches is a hard task and time-consuming. Therefore, methods and algorithms that offer to classify and predict these data sets have become great allies to understand better the huge amount of data and information. Among the algorithms used to classify and predict large data sets, supervised and unsupervised learning algorithms have been gained highlights in the last years. Supervised learning algorithms are algorithms that initially need to be taught by a set of labeled data, whereas the unsupervised learning algorithms do not need to be taught, and are able to work with unlabeled data. Support Vector Machine (SVM), Neural Networks, and K-means are supervised and unsupervised learning algorithms that through robust mathematical calculations are responsible for auxiliary in the classification and prediction of data sets. The work used K-means, Support Vector Machine, and Neural Networks to classify and predict different electrofacies in a data set which were provided eight input features, they are, Gamma ray, Neutron, Resistivity, Water Saturation, Density, Porosity, P-velocity, and S-velocity. Between the supervised learning algorithms used in the research (SVM and Neural Networks), it is clear that according to statistical validation of models, confusion matrix and K-fold cross validating, the Neural Network presented a better performance at all scenarios in relation to the Support Vector Machine. The accuracy evaluation of SVM algorithm using, confusion matrix and K-fold cross-validating was 92.9% and 86.6% respectively. While the accuracy evaluation of the Neural Networks algorithm was 96.6% and 96,9% respectively. The better performance can be verified in the others parameters as, precision, recall, and F1 score, all parameters responsible by evaluate the model performance. The only point where the Support Vector Machine has significant advantages

over Neural Networks is at execution time. While the model based on SVM is capable of predicting the electrofacies in 0.096s, the model based on Neural Networks takes 41.45s to complete its predictions. In other words, the model based on SVM is about 432 times faster than the model based on Neural Networks. The data set of the well log predicted had two electrofacies predicted in the supervised learning algorithms, showing convergence between the models. The models applied in the research with AI algorithms proved to be robust for classifying and predicting electrofacies and a promising method capable of automating the processing of electrofacies classification in Brazilian pre-salt, and easily adjustable to work with high accuracy in other regions. The research shows new ways to classify and predict electrofacies from well logs with more agility, reliability and accuracy, and tries to open up new possibilities for Artificial Intelligence algorithms in the world of geosciences.

**Keywords**: Artificial Intelligence; Electrofacies; Carbonates; Brazilian Pre-salt

# List of Figures

# List of Tables

# Contents

# Introduction

In the last years, Artificial intelligence (AI) algorithms have been inserted in several areas of knowledge and the oil and gas (O&G) sector has not been left behind. In recent years the O&G has made AI algorithms more present in the industry and literature. A lot of works trying to apply AI algorithms can be found in the recent literature ([KVLD12], [SAKA12]), however, due to be a new field of study, there are many challenges to be overcome. The introduction of them in petroleum industry has become necessary because the amount of data collected nowadays has increased at high rates and, for processing these huge amounts of data are necessary robust mathematical algorithms. Analyzing a large volume of data is required in order to develop a comprehensive understanding of reservoir distributions and their production performance characteristics [EDFK10]. Among the AI algorithms there are two major groups, the supervised Learning algorithms [KZP07] and the unsupervised Learning algorithms [Jai10]. The current research developed three different models, two models using supervised learning algorithms and one model applying unsupervised learning algorithm.

The unsupervised learning algorithm that was used to classify the data set it was the k-means describe by Mac Queen in [Mac67] and Sariel and Bardia in [HPS05] . Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning, and cluster analysis is the formal study of methods and algorithms for grouping, objects according to measured or perceived intrinsic characteristics or similarity [Jai10].

The supervised learning algorithms that were used to classify and predict the electrofacies in the research were, Support vector machine (SVM) [Vap13] and Neural Networks [Spe91][GG93][AB09]. One of the most common applications of SVM and Neural Networks algorithms is to recognize and predict patterns in different data set and in this way, the SVM [TB20] and Neural networks have been widely used to solve complex real-world problems [XTJ17].

Much progress has been made to understand and improve learning algorithms, but the challenge of artificial intelligence (AI) remains [Ben09].

In geology of petroleum, an indirect way to identify the rock properties in the subsurface is through the geophysical well-logs, which are responsible for indirectly measuring the rock properties. These geophysical well logs are able to identify the porosity, density, resistivity, and radioactivity in the reservoir [TB20]. From these indirect measures of the rocks in subsurface, the Geologists are capable to recognize patterns in different packs of the rocks. For the patterns found by the expert's Geologists from the well logs are given the name of electrofacies [SA80], thus, the term electrofacies are numerical combinations of petrophysical log responses that reflect specific physical and compositional characteristics of a rock interval[Dav18]. Traditionally electrofacies have been identified manually with the aid of graphical techniques like crossplotting from wire-line logs [KK06a]. But people are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Most recently several AI algorithms have been introduced to try automating the task of electrofacies identification [KK06b]. Electrofacies have been used widely in petroleum prospecting and reservoir characterization as a tried to distinguish different beds in a petroleum field as well as in the correlation with the lithofacies [Nic09], [Mia10].

The current research is development from carbonates reservoirs of the Brazilian pre-salt, located in Santos basin, *Gato do Mato oil field*. The Brazilian pre-salt is a province characterized by carbonate reservoirs, microbial and coquina rocks, buried at a depth that surpasses 5.000 meters, distributed in the Santos and Campos sedimentary basins, located at the southeast Brazilian coast [JM16]. Brazilian pre-salt carbonates reservoirs are derived from lacustrine environments, and appears to be laterally continuous over tens of kilometers [CWG08]. The laterally continuity of them also suggest the lateral continuity of electrofacies, and in this case, the interpolation of electrofacies among the well logs, in the oil field, can be done with a high level of safety.

The Santos Basin is localized in the southeast of Brazil, with an area approximate of 350.000 $km^2$ and the sediment thickness in some areas is higher than 10 km [CAC+08]. The figure 2.4 presents the Pre-salt distribution, as well as the study area. Estimates in Santos Basin suggest that the potential volume of oil reserves is higher than 100 billion

barrels [Sau16], which would position Brazil as having the fifth-biggest world reserves. The stratigraphic section studied in the Santos Basin is localized under the evaporitic unit formed during the post rift phase in the Aptian Stage. The main stratigraphic units that compose the study area are Itapema and Piçarras Formations. The Itapema is located immediately below the evaporitic section and was formed between the Late Barremian to Early Aptian. Based on the paleogeographic distribution, in the distal portions, the Itapema Formation was formed by marine incursions that were responsible to deposited dark shales and carbonate rocks [Ara14], while the proximal portions are constituted by conglomerates and sandstones deposited by alluvial fan [MMGM07]. The Piçarras Formation corresponds to those alluvial fan sediments, composed of conglomerates and lithic sandstones deposited during the Barremian Stage. Volcanic rocks of the Camboriú Formation (Upper Necomian) constitute the basement of the basin [MMGM07].

In order to identify different electrofacies related to carbonate reservoir rocks, the work proposes to use three different models based in the algorithms, k-means, SVM, and Neural Networks for classifying and predicting the electrofacies of the *Gato do Mato* Oil Field. The study area was chosen due to its strategic location inside the Santos basin, and because the lack of information about the *Gato do Mato* Oil Field in literature. Therefore the current research has with one of the goals to become more understandable and accessible the geological information about the *Gato do Mato* Oil Field.

The current research also has been contributed to the discussion about the importance and value of AI algorithms in electrofacies classification and prediction. Overcoming the challenges involving this process could result in more precision and agility in electrofacies classification. Therefore, due to the high impact to literature and Petroleum Industry it is crucial to make more visible the theme AI algorithms in O&G. The next step to the future is taken every day with the advance of knowledge in different areas. At the moment, it is unquestionable that AI algorithms are the bridge to the faster development of many areas of the knowledge. Thus, it makes clear that AI algorithms can be a great partner in this long journey to the future.

# Chapter 1

# General Considerations

The chapter will go to lead with introductory informations about the motivations, objectives and justifications associated to the research developmented in the last two years. The section 1.1 shows how the final thesis structure was thought and constructed. From section 1.2 the reader will know the motivations and justifications that induced the authors to carry out the current research. In section 1.3 are described the main goals that the research aimed to achieve.

## 1.1 Thesis Structure

The thesis was wrote in the book format in order to become the organization of information more accessible and clear to the reader. The chapters were organized and grouped according to the subject of the topics. Topics with similar contents were grouped in the same chapter, and topics with different subjects were separated according to the most similar chapter. The thesis is composed of six chapters and three appendixes. The chapter 1 contains the general considerations about the thesis. Section 1.1 contains a brief explanation about the thesis structure. In section 1.2 a brief description of the problem it is presented. Section 1.3 it is responsible to talk about the justifications that make the research to happen. And in section 1.4 are presented the objectives. The chapter 2 talk about the geology of the study area, in the section 2.1 the geological context it is presented to the reader, here in this section is possible to understand a little more of the evolution of the Santos basin, the Formation presents in the area, and the lithology expected to the area. The section 2.2 shows the location of the study area. The chapter 3 is responsible

for demonstrating the theoretical foundation of AI algorithms and geophysical well logs. In the section 3.1 it is expose the algorithms, K-means, Support vector machine and Neural Networks. The section 3.2 it is presents the geophysical well logs. The chapter 4 contains the part of materials and methods. The section 4.1 shows the data set utilized in the research and how it was created. In the section 4.2 it is explained the data science stage. The section 4.3 presents how was made the data set division in training and test data set. In section 4.4 the working flowchart of the models applied in the research are illustrated. Section 4.5 demonstrates the statistical validation used for evaluating the models. And the last section 4.6 of the chapter show the vertical heterogeneity method. The chapter 5 shows the results obtained in the research and three appendixes(A, B, and C. The last chapter 6 talk about the final considerations, in the section 6.1 it is discussed the results obtained and the section 6.2 are presents the main conclusions of the research. The appendices section is composed of articles that were prepared with the research results. The articles presented in appendix A, B and C shows the evolution of the research that begins working with unsupervised machine learning (K-means, appendix A), supervised machine learning (Support Vector Machine, appendix B), and ends using in Neural Networks algorithms (appendix C). The thesis structure was thought to be a simple and objective file. A file that, the reader, will be able to read all the content in just one go. From this main idea, the results of the thesis were presented by articles that derived from the current research. This thesis format is closer to author preferences than the traditional thesis format. The results, chapter 5, were referenced to the Appendix A, B and C.

## 1.2 Problem Description

The knowledge about a petroleum field it is commonly obtained of two ways, by interpretations of seismic surveys, and geophysical well logs. The seismic data has the characteristics of describe the macro features in a petroleum oil field, while the geophysical well logs are capable of reveal micro aspects of rock and fluids composition. A way to correlate the geophysical well logs with rock and fluids characteristics in subsurface is to identify and associates the geophysical well logs behaviors with samples of rock and fluids obtained of the well. Another way to correlate the geophysical well logs with

rock and fluids characteristics is to verify the papers of literature that describe behaviors of geophysical well logs for different types of rock and fluids [Tit12], [ES07], [KMP05], [AKG04], [WDW+15]. From the process of associating characteristics of rock and fluids with geophysical well logs it is possible to expand and find the electrofacies [KK06a], [KVLD12].

The classification of electrofacies is a task usually realized by geologists due to their geological and geophysical knowledge. The conventional way is to make a manual electrofacies classification according to the patterns found in the geophysical well logs. These electrofacies are in many times responsible to help the understanding of rocks in the subsurface and in the delimitation of the reservoir and non-reservoir zones. But, often, the manual classification of electrofacies can become a problem. The first problem is related to a non-standardization in the electrofacies classification. In other words, it is possible to have different classification patterns according to each geologist. The second problem is related to the number of geophysical well logs present in an oil field. The volume of data is higher in oil fields with a large number of wells, consequently, the manual classification can become expensive and time-consuming. In addition, the classification can take the risk of not being standardized. Therefore, the research trying to solve this two problems related with the electrofacies classification.

## 1.3 Justifications

Due to the strategic location and proximity to large pre-salt productive areas of hydrocarbons, the Campo Gato do Mato Oil Field reveals to be an area with high hydrocarbon production potential.

The Gato do Mato is an oilfield present in Brazilian Pre-salt[MMGM07]. It is composed of carbonate rocks from the Pre-salt section, being an ideal oilfield to build models of classification and prediction of electrofacies from geophysical well logs of carbonate rocks.

The study area, Gato do Mato Oil Field, is a poorly explored region compared to adjacent areas such as the Lula Oil Field, thus, one of the reasons for the current research to be developed in the Gato do Mato is to make the geological knowledge of the region more accessible. Once a time that the interpretation of the association among the

geophysical well logs reveals some physical properties as porosity, rock matrix, and fluids, consequently, the electrofacies classification makes more accessible the physical properties of the carbonates rocks in the subsurface of the study area to the public.

As pioneer research about Gato do Mato Oil Field, the research brings AI algorithms in geophysical well logs, and adds new results, reservoir zones are present in the study area with thickness that can reach 40 meters, but due to the complexity and the high compositional variation of carbonates rocks presents in the study area, the reservoir zones have in general many heterogeneities, and have many non reservoir zones intercalated between the interval of reservoir zones. And perspectives, the vertical heterogeneity found in Pre-salt zones of the study area reveals possibles futures actions to be done in an eventual production phase, such as, the choice between vertical, horizontal and directional wells, in order to maximize the hydrocarbon production of the Oil field. In some reservoirs the horizontal well may improve drainage by increasing the area of the wellbore in contact with the reservoir[WJW90]. In reservoirs with high vertical heterogeneity, such as the study area, many times the use of horizontal and directional wells is often advisable in order to maximize the hydrocarbon extraction.

The work also provides new techniques, models based on AI algorithms to predict and classify patterns of geophysical well logs, to the petroleum industry in the reservoir characterization context. With the automation of the prediction and classification of electrofacies, and the delimitation of the reservoir and non-reservoir zones, the work in the oil field gains more agility, constancy, and precision. Thus, the work contributes to a better Geological and Geophysical understanding of the Gato do Mato Oil field once that explores geophysical well logs and provides Geological information about reservoir and non-reservoir zones in the oil field. In addition, the research also contributes to models based on algorithms of Support vector machine and Neural Networks to automate the process of classification and prediction of electrofacies.

## 1.4 Objectives

The research was developed with the main goal of creating a model capable to classify and predict electrofacies from geophysical well logs in an automated way. The area chosen to create the model was the Gato do mato oilfield. From the wells presents in the Gato

do mato oilfield were chosen two wells to create and test the model, 1-SHEL-23-RJS and and 1-SHEL-26-RJS. It is hoped that the automated classification and prediction of electrofacies helps to save time and auxiliary in the understanding of the distribution of reservoir and non-reservoir zones in the study area.

To achieve the main goal, some secondary goals(targets) had to be reached before. Targets are steps with less complexity to be reached in order to achieve the main goal. Divide and conquer is often used in complex algorithms problems in order to divide a complex problem into smaller pieces, easier to solve. The targets of the research were, the creating of a data set able to represent the three characteristics of a rock formation, classifying electrofacies able to distinguish zones reservoir and non reservoir, creating models with different AI algorithms to compare the models performance, and obtaining statistical tools able to evaluate models performance. The main goal and targets are listed below:

- Creating of a model capable of classifying and predicting with high accuracy and reliability electrofacies from geophysical well logs in an automated way.
  - Creating of a data set from geophysical well logs capable to represent the three rock formation characteristics (rock matrix, fluids, porosity) and that be present in the all wells of the Gato do Mato;
  - Classifying electrofacies able to distinguish zones reservoir non reservoir. This is important because allows to correlate the electrofacies with zones reservoir and non reservoir;
  - Creating models with different AI algorithms to compare the performance between them;
  - Obtaining statistical tools able to evaluate the performance of models with precision and reliability.

# Chapter 2

# Geology of the Study Area

This chapter presents the Regional Geology of the Santos Basin with the goal of introducing to the reader a brief context of the study area. The chapter also presents the location of the study area and the relation between the Santos Basin and other basins present in the southeast margin of the Brazil. Section 2.1 introduces the regional geology of the Santos Basin with illustrations about the Stratigraphic chart and considerations about the Formations that compound the Basin. In section 2.2 the reader will know the location of the study area, Gato do Mato oil field, in relation to the Brazilian Pre-salt and other oil fields presents in the region.

## 2.1 Geological Context

Estimates in Santos Basin suggest that the potencial volume of oil reserves is higher than 100 billion barrels [Sau16], would position Brazil having the fifth biggest world reserves. The confirmation of the exploratory viability of the hydrocarbon reservoirs below of the evaporitic sequence in the Parati Oil field, in 2005, made the Santos basin the main receptive of investments in exploration and production by Petrobras, guaranteeing the beginning of a new exploratory and productive cycle [SS16].

Another important aspect related to the attraction of investments for implementation of the production units in the Santos Basin was the fact that its geographic positioning to be adjacent to the Campos Basin, which initially allowed the reduction of the implementation costs of the necessary infrastructure to the development of the works [SCS19].

The potential volume of reserves, combined with the big accumulations of light oil

[Sau16], the high productivity of producers wells, many of them overcoming the production of the big part of the producers basins of hydrocarbon in the country, allows the progressive reduction of the production costs of the petroleum barrel, establishing a competitive production in relation to the others producers regions of the planet [NG17]

The Santos basin presents a very complex stratigraphic framework (Figure 2.1 and 2.2), this is a result of the over position of several geological events. These events are printed in the framework stratigraphic of the basin, but its expression is unequal, in terms of magnitude and representativeness, in different structural sectors of the basin[CAC+08]. As observed by [SS16], "the Santos Basin has two important proprieties in relation to the others pre-salt basin": (i) absence of emerged portion of the basin [Gar12], and (ii) occurrence of magmatism forming the basement in the basin, represented by the Camboriú Formation, of the Superior Neocomian [MMGM07] similar to what happens in Campos Basin, represented by Cabiúnas Formation, also from Superior Neocomian [TFMA08].

As suggest [MMGM07], the stratigraphy of Santos Basin (Figure 2.1 and 2.2) can be better understood , dividing the basin into three stages, rift, post-rift, and drift. Thus, in the Santos basin:

The drift stage was deposited from the Albian until the recent, and it is characterized by the Itamambuca Group composed of the Marambaia Formation. The Fraude Group represented by Santos, Juréia, and Itajaí-Açu Formations. The Camburi Group, composed by Florianópolis, Guarujá, and Itanhaém Formations. The depositional environments that characterize this stage are coastal, platform, slope, and deep.

The Marambaia Formation is related to turbidite flows and is composed of pelitic sediments interpolated with sandstones from Maresias Member.

The Santos Formation is derived from an alluvial fan and is represented by red conglomeratics sediments. The Juréia Formation also derived from an alluvial fan and is represented by red arenaceous and pelitics sediments. Both Formations represents a regressive phase in the Santos Basin [MMGM07].

The Itanhaém Formation is formed by the turbidite and hyperpycnal flows, its lithology is composed of pelitic sediments interpolated with sandstones from Tombo Member. The Itanhaém Formation represents a transgressive phase [MMGM07].

The post-rift and rift stage is composed by the Guaratiba Group and corresponds to the interval of 140-110 Ma. The depositional environments where the Guaratiba Group

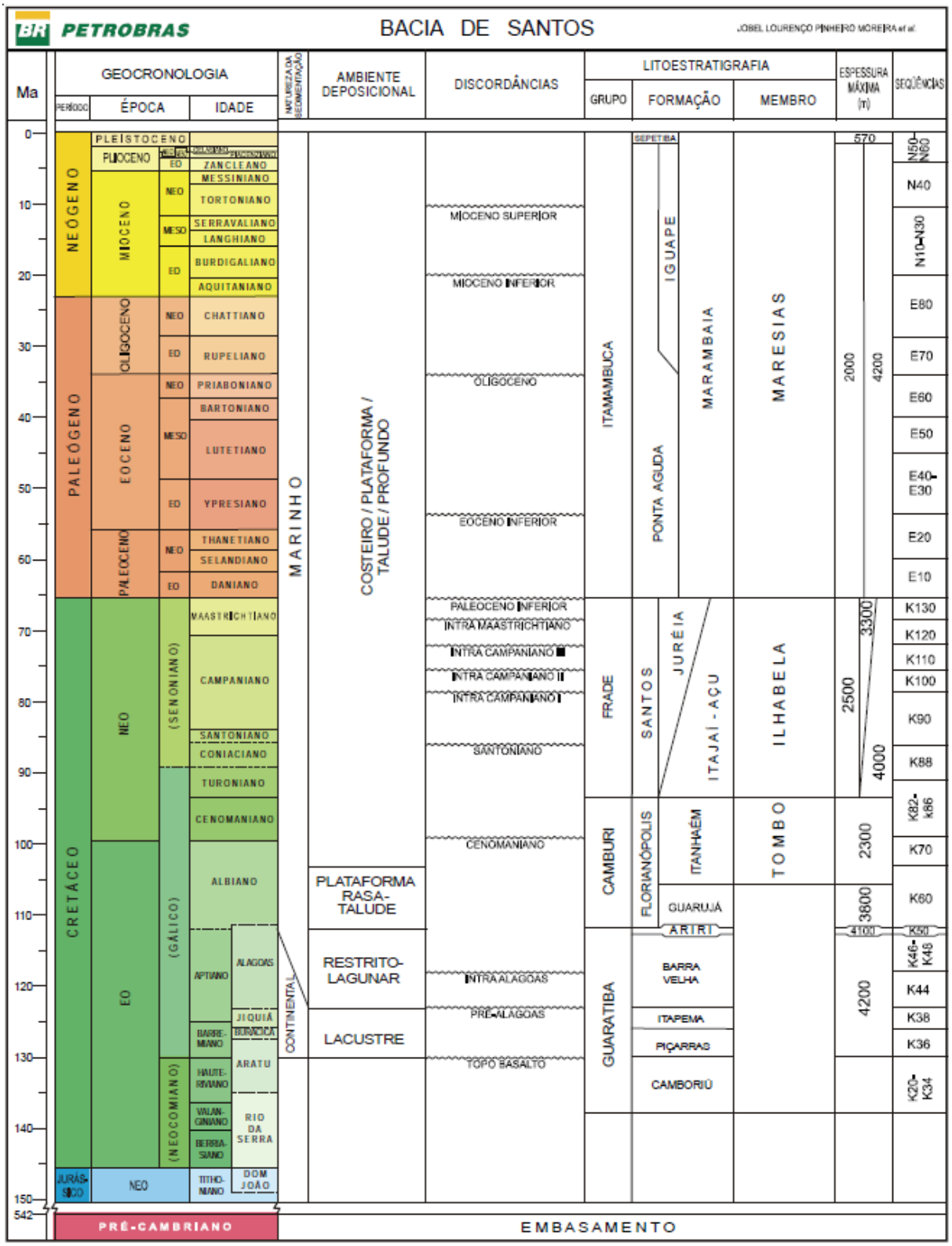



Figure 2.1: Stratigraphic chart from Santos basin part I [MMGM07].

developed were, restrict lagoons and lacustrine.

The post-rift stage was deposited between Aptiano and the beginning of Albiano, and

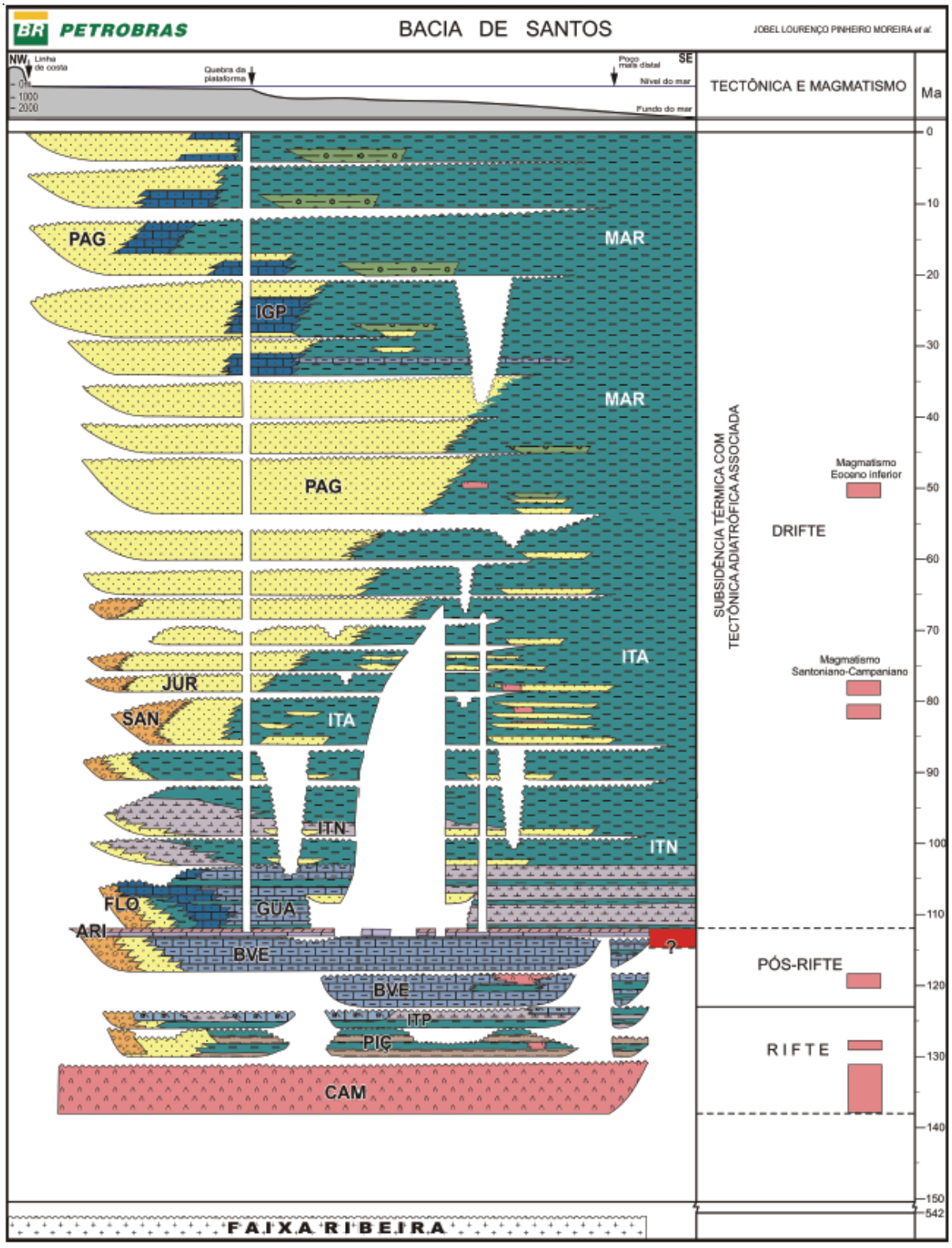



Figure 2.2: Stratigraphic chart from Santos basin part II [MMGM07].

it is characterized by the Barra Velha and Ariri Formations. The Barra Velha Formation is formed by sediments deposited in the initial stage of the thermal subsidence. Its lithology

is composed by limestone stromatolitic and carbonate shales.

The rift stage extends from the Hauterivian to the Aptian and it is characterized by the Camboriu, Piçarras, and Itapema Formations. The Itapema Formation is formed by sediments from the final of the half-graben formation process. The lithology is composed by grainstones, wackestones and packstones bioclastics, and carbonate shales. There is a significant lateral change in the facies [MMGM07]. The Piçarras Formation is composed by sediments deposited in the maximum activity of the half-graben formations. Its lithology is formed of sandstones and dark shales, rich in organic matter.

The figures 2.1 and 2.2, shown the Chronostratigraphic chart of the Santos Basin according to [MMGM07]. In the Chronostratigraphic chart is possible to observe the Geochronology, Depositional environments, Lithostratigraphy, and the phases of tectonism and magmatism in the Santos Basin.

The stratigraphic section studied in the Santos Basin is localized under the evaporitic unit formed during the post rift phase in the Aptian Stage. The main stratigraphic units that compose the study area are Itapema and Piçarras Formations. The Itapema is located immediately bellow the evaporitic section and was formed between the Late Barremian to Early Aptian. Based on the paleogeographic distribution, in the distal portions, the Itapema Formation were formed by marine incursions that were responsible to deposited dark shales and carbonate rocks [Ara14], while the proximal portions are constituted by conglomerates and sandstones deposited by alluvial fans [MMGM07]. The Piçarras Formation correspond to those alluvial fans sediments, composed by conglomerates and lithic sandstones deposited during the Barremian Stage. Volcanic rocks of the Camboriú Formation(Upper Necomian) constitute the basement of the basin [MMGM07].

## 2.2 Location of the Study area

The Santos Basin is localized on the southeast margin of Brazil, with an area of approximately 350.000km$^2$, the sediment thickness in some areas is higher than 10 km [CAC+08]. The Santos Basin has developed from a rifting process of the Gondwana Paleocontinent, this event has begun in the Eocretaceous, the basin is present in the margin of the states Rio de Janeiro, São Paulo, Paraná, and Santa Catarina represents one of the greatest depressions of the Brazilian continental margin, and contrary to the

Campos Basin has the whole area submerse. The Santos Basin is bordered by the Campos Basin in north and the Pelotas Basin to the south, to the west by Serra do Mar and to the east by the São Paulo plateau [GMDS+08](figure 2.3). In the Santos Basin, the oils generate has origin saline from rocks deposited in lacustrine environments during the Aptian stage and it is represented by the Guaratiba Group. As illustrated in the figure 2.3, the pre-salt reservoir has the biggest portion inside the Santos basin, thus, is of extremely importance the studies about the Basin.


Figure 2.3: Santos basin location in relation to the Campos and Pelotas Basin [RST12].

The study area, Gato do Mato oil field, is located in the southwest of the brazilian pre-salt and in the middle of the Santos basin. The Gato do Mato location can be visualized in the figure 2.4, according to the figure 2.4 it is possible to observe that the study area is closer to the coastline than Lula oil field, what in many cases can facility the logistic

of the hydrocarbon transportation. The study used information over two well-logs from Pre-Salt reservoir, with more than 5.000 meters of depth. The interval of interest has up to 400 meter and presents geophysics and lithologicals characteristics of the carbonate rocks, and was used to the classification and prediction of electrofacies by AI algorithms.

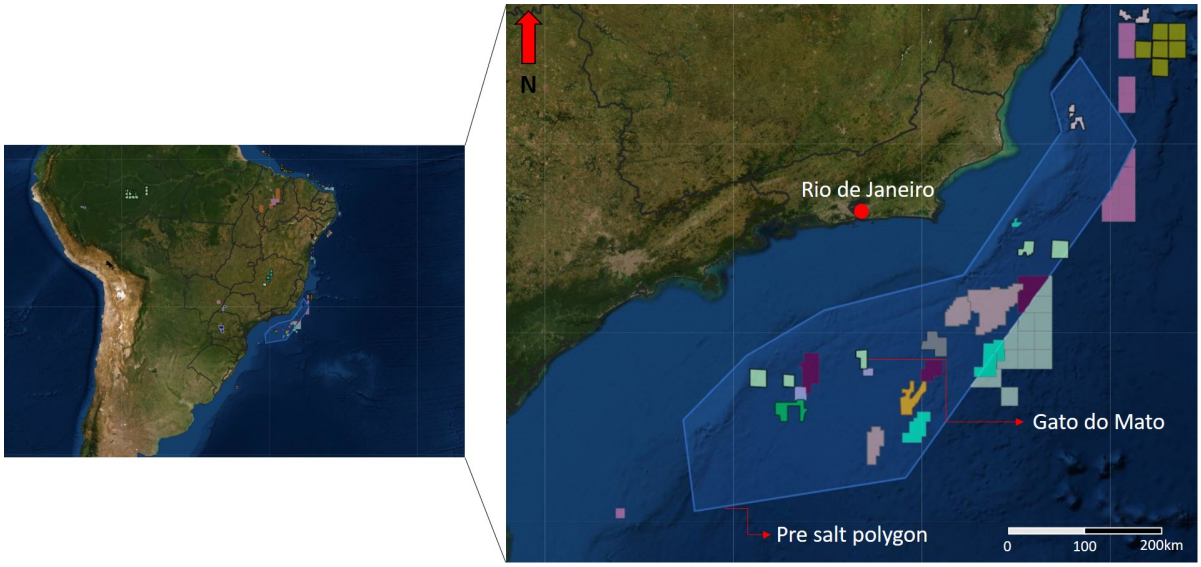



Figure 2.4: Location of the study area, Gato do Mato oil field, in relation with Brazilian pre-salt.

# Chapter 3

# Theoretical Fundamentals

The chapter will present the theory behind the algorithms applied in the research to build the models, and the theory about Geophysical well logs used to compose the data set of the research. Section 3.1 will provide a brief theoretical fundamentals of AI algorithms applied in the research, and show the two main groups presents that compound the AI algorithms they are, (i) supervised learning algorithms, and (ii) unsupervised learning algorithms. From section 3.2 the reader can find out the theory about the Geophysical well logs.

## 3.1 AI Algorithms

Artificial intelligence is the term used to describe the solving of problems and make decisions by machines similar to the human mind. In the literature ([Cop04], [HTF09], [STS16a]), there are a huge amount of Artificial Intelligence algorithms, these algorithms sharing in common mathematical algorithms able to recognize and solve problems with high precision, accuracy, and agility. In general they can be subdivided in two main groups, (i) supervised learning algorithms, and (ii) unsupervised learning algorithms. The section 3.1.1 explains how the unsupervised learning algorithms working, while in the section 3.1.2 is possible to find the explanation about the supervised learning algorithms.

### 3.1.1 Unsupervised Learning

In unsupervised learning algorithms, there is not a labeled data set, so the algorithms need to learn for themselves. With the proliferation of massive amounts of unlabeled

data, unsupervised learning algorithms–which can automatically discover interesting and useful patterns in such data–have gained popularity among researchers and practitioners[CA16].The goal of unsupervised learning is to directly infer the properties of a data set unlabeled without the help of a supervisor or teacher providing correct answers or degree-of-error for each observation[HTF09]. Unsupervised classification tasks the labels are not provided, and the task of the algorithm is to find a "good" partition of the data into clusters[Gha03]. Within unsupervised learning, clustering is probably the most popular technique, because it is responsible for grouping similar data with significant success. Clustering is a method that unsupervised learning algorithms have to identify groups of similar objects in multivariate data sets without having any prior knowledge of their group memberships. Given a data set, a clustering algorithm will classify each data point into a specific group. Each cluster that arises during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters[RM17].

#### 3.1.1.1 K-means

K-means clustering is a method commonly used to automatically partition a data set into k groups [WCRS01].The data clustering, also known as cluster analysis, try to discover the natural grouping(s) of a set of patterns, points, or objects [Jai10]. The aim of cluster analysis is to classify a data set into groups that are internally homogeneous and externally isolated on the basis of a measure of similarity or dissimilarity between groups [KSIN15].

Several clusters algorithms have been proposed to try classify different data set, but due to its simplicity, the K-means algorithm have been the most commonly used in the literature. Given a set of $n$ data points in real $d-dimensional$ space, $\mathbf{R}^d$, and an integer $k$, the problem is to determine a set of $k$ points in $\mathbf{R}^d$, called centers, so as to minimize the mean squared distance from each data point to its nearest center [KMN+02].

Let $X = \{x_i\}, i = 1, ..., n$ be the set of n d-dimensional points to be clustered into a set of K clusters, $C = \{c_k,\ K = 1, ..., k\}$. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized [Jai10]. The K-means works to minimize the sum of the squared error (Eq.3.1.1). Minimizing this objective function is known to be an NP-hard problem

(even for K = 2) [DFK+99]. K-means is better explained in the steps (I) to (V) and by the work flowchart illustrated in the Fig.3.1

(I) Assigning the centroids;

(II) The distance between each point to the centroid is calculated. N-points and K-centroids;

$$N \times K = number\ of\ distances\ calculated$$

(III) Each point is putting in the class according to the centroid distance. The point is embody by the nearest centroid and will belong to the class represented by centroid;

(IV) New centroids are calculated for each class and the value of centroid coordinate are refined. For each class that has more than one point, the new centroid coordinate is calculated using coordinate average of all points belongs to the class;

(V) The algorithm repeat the third and fourth step repeatedly until the convergence. When in the loop n the centroid coordinate doesn't change in relation to the previous loop (n-1), then the process finish and the centroid coordinate is found.

The figure 3.1 shows the K-means work flowchart.

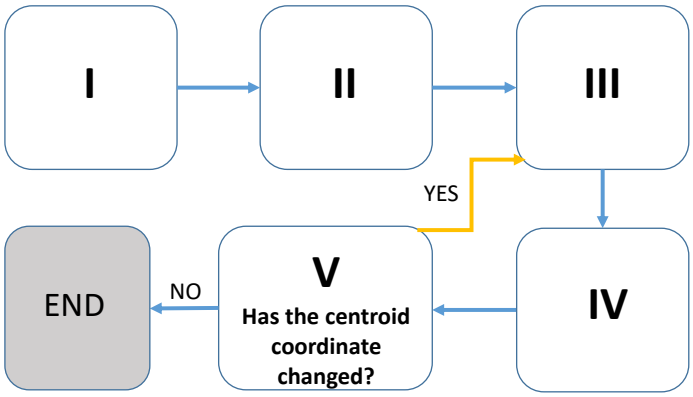



Figure 3.1: Work flowchart of the K-Means algorithm.

In the K-means algorithm the users need to provide the number of classes that fulfill their wishes, but the wrong choose of the number of clusters will result in K-means clustering algorithm with high errors and poor's cluster results. Thus, the Elbow Method can be an important and complementary method to help in the choose of the correct number of clusters in a dataset.

#### 3.1.1.2 Elbow Method

Elbow method is a method which looks at the Sum Square error percentage of variance explained as a function of the number of clusters [BA14]. The Elbow method is expressed by

$$SSE = \sum_{K=1}^{K} \sum_{Xi \in Sk} \| Xi - Ck \|_2^2 \tag{3.1.1}$$

Where SSE is the sum of the average Euclidean Distance of each point against the centroid [MHW18].The letter K is the number of clusters, Xi is the data present in each cluster, Ck is the K-th cluster and Sk is the set of points inside the Ck cluster.

The Sum Square error explained by the cluster is plotted against the number of clusters. The first cluster will add much information, but at some point the marginal gain will drop dramatically and gives an angle in the graph [BA14]. When the increase of the cluster number does not varies considerably the SSE, then the best K value was found.

Most of clustering algorithms are designed only to investigate the inherited grouping or partition of data objects according to a known number of clusters. Thus, identifying the number of clusters is an important task for any clustering problem [KM13].

### 3.1.2 Supervised Learning

Supervised learning is the construction of algorithms that are able to produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances[STS16a]. Supervised learning is the name given for a set of algorithms responsible for learning to categorize data from a data set labeled and, then, to apply the knowledge learned to unlabeled data sets. In supervised tasks the training data consists of training patterns, as well as their required labeling [Gha03]. A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples, thus the main goal of the supervised learning is to approximate new input data so well that you can predict the output variables for that data. Learning stops when the algorithm achieves an acceptable level of performance. We can say that supervised learning is similar to say "learning with a teacher". With supervised learning, there is a clear measure of success or lack thereof, which can be used to judge adequacy in particular situations and to compare the effectiveness of different methods over various situations [HTF09].There are two ways to work with supervised

learning problems, (a) the first way is work with the resolution of problems that involved prediction of a specific target, while the (b) second are problems that the data set need to be classified.

#### 3.1.2.1 Support Vector Machine

The Support Vector Machine is an algorithm to predict and classify that was developed by Vladimir Vapnik in 1960s, and in the nineties the algorithm was highlighted to solving pattern recognition problems through of the Vapnik's works ([Vap95]; [Vap98]). Recently, SVM has been considered one of the top methods in pattern classification. It has many attractive properties such as a strong mathematical foundation, few tuning parameters, fast classification and high generalization capability [FA19]. The SVM method maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space. The algorithm works with a concept of maximizing the minimum distance from hyperplane to the nearest sample point [STS16b]. This basically involves solving a quadratic programming problem [SV99]. The technique of SVM was first developed for the restricted case of separating training data without errors, later was enhanced the case of separating data with errors. Thus, there are two historic Support Vector Machine algorithm, (i) the hard margin Support Vector Machine and (ii) the soft margin Support Vector Machine that will allow for an analytic treatment of learning with errors on the training set [CV95]. In both algorithms the elements, hyperplane, margin and support vectors are essential. The hyperplane is understood as a plane that separate the data, the margin is the distance of the hyperplane and the support vectors and the support vectors are the closest points of the hyperplane. The optimal hyperplane is defined as the linear decision function with maximal margin between the vectors of the two classes [CV95]. According to [CV95] finding the optimal hyperplane in the hard margin case the algorithm considers the set of labeled training patterns:

$$(y_1, x_1), ..., (y_l, x_l) \quad y_i \in -1, 1 \tag{3.1.2}$$

The data set is linearly separable if there is a vector W and a scalar b such that

$$y_i(W \cdot x_i + b) \geq 1 \quad i = 1, ..., L \tag{3.1.3}$$

are valid for all elements of the training set. The optimal hyperplane is given by

$$W_0 \cdot x + b \tag{3.1.4}$$

The maximal distance between the projections of the training vectors of two different classes under the constraints (3.1.3) will ensure an optimal hyperplane (3.1.4)

$$\varphi = W \cdot W \tag{3.1.5}$$

for this the equation (3.1.5) has to be minimized. To do this a standard optimization technique is used and a Lagrangian is constructed

$$L(W, b, \Lambda) = \frac{1}{2}(W \cdot W) - \sum_{i=1}^{L} \alpha_i [y_i(x_i \cdot W + b) - 1] \tag{3.1.6}$$

where $\Lambda^T = (\alpha_1, ..., \alpha_l)$ is the vector of non-negative Lagrange multipliers corresponding to the constraints (3.1.3). To find the hyperplane in the soft margin case the reasoning is similar and can be found in the article of [CV95].

The SVM algorithm deals very well with linear data sets, in cases where a straight line is sufficient to separate a data set, but for problems with non linear data sets is not possible to find a straight line to classify the data set, in this case the Kernel Trick [BGV] is used. The basic idea is that if a data set is inseparable in the current dimensions, the kernel trick will carry the input data set to a higher dimensional space and by choosing an adequate dimension, the data set points become linearly separable or mostly linearly separable in the high-dimensional space [AW99]. Thus, the SVM with the Kernel trick [HSS08] currently has been ensured the solving of linear and non linear problems efficiently. Beyond that references cited before it is recommended the articles [Bis06] and [SS04] as good references for the theory and practicalities of SVM.

The Support Vector Machine algorithm implemented in the scikit-learn Python library, version 0.24.2, have been improved in relation with the initial implementation of the algorithm made in the articles [Vap95], [Vap98]. Internally the library uses [CL11] and [FCH+08] as guidelines for implementation of the Support Vector Machine algorithm.

In the model I (SVM), the Kernel parameter that better adjusted to the data set was the "Linear", thus the Kernel was set with, Kernel = Linear, the article [CS01] shows

how the multiclass Kernel algorithm is implemented in the Python library, version 0.24.2. The Python version used to implement the model in the research was 3.8.3.

The hyperparameters used in the SVM implementation by Scikit-learn [PVG+11] Python library are, $C$ (the penalty of the error), $Kernelfunction$ (the Kernel functions available in the library are, 'linear','rbf', 'poly' and 'sigmoid'), *degree* (available to the polynomial Kernel function. Ignored by all other kernels), *gamma* (it is the Kernel coefficient for 'rbf', 'poly' and 'sigmoid' Kernel functions), $coef0$ (Independent term in kernel function. It is only significant in 'poly' and 'sigmoid'), *shrinking* (it is a parameter that auxiliary the convergence velocity of the algorithm, decreasing the training time), *probability* (when enabled, probability estimates are calculated), *tol* (it is a stopping criterion), *cache_size* (the size of the kernel cache in MB), *class_weight* (set the parameter $C$ of class i. If not given, all classes are supposed to have weight one), *verbose* (when enabled return a verbose output), *max_iter* (number maximum of interactions), $decision\_function\_shape$(it is the decision function present in the classifier), *break_ties* (if true, it is responsible for breaking ties according to the confidence values of the decision function), *random_state* (it is responsible for controls the pseudo-random number generation for shuffling the data for probability estimates). In the Scikit-learn [PVG+11] Python library, version 0.24.2, there are two stopping criterion for the SVM algorithm, the first it is the number of maximum interactions (*max_iter*) that the algorithm will made in the training data set, in this setup the algorithm will stop when the maximum interactions number was achieved. The second stopping criterion it is the tolerance factor (*tol*), in this case, when the loss function or the validation score is not improving by at least a tolerance factor (*tol*) for two consecutive interactions, convergence is considered to be reached and training stops. In the research was utilized the tolerance factor with stopping criteria.

The model applied in the research has setup the hyperparameters as follow, $C = 100$ (the penalty of the error), $Kernelfunction = linear$ (the Kernel functions available in the library are, 'linear','rbf', 'poly' and 'sigmoid'), $degree = 3$ (Ignored by linear kernel function), $gamma = scale$ (it is the Kernel coefficient for 'rbf', 'poly' and 'sigmoid' Kernel functions. Ignored by linear kernel function), $coef0 = 0.0$ (Independent term in kernel function. Insignificant in linear kernel function), $shrinking = true$ (it is a parameter that auxiliary the convergence velocity of the algorithm, decreasing

the training time), $probability = false$ (when enabled, probability estimates are calculated), $tol = 0.001$ (it is a stopping criterion), $cache\_size = 200$ (the size of the kernel cache in MB), $class\_weight = none$ (set the parameter $C$ of class i. If not given, all classes are supposed to have weight one), $verbose = false$ (when enabled return a verbose output), $max\_iter = -1$ (number maximum of interactions, -1 for no limit), $decision\_function\_shape = ovr$ (it is the decision function present in the classifier), $break\_ties = true$ (if true, it is responsible for breaking ties according to the confidence values of the decision function), $random\_state = 3$ (it is responsible for controls the pseudo-random number generation for shuffling the data for probability estimates).

Support Vector Machine (SVM) has been proven to perform much better when dealing with high dimensional datasets and numerical features. Although SVM works well with default value, the performance of SVM can be improved significantly using parameter optimization [SPBW16]. However, parameter optimization can be very time consuming if done manually especially when the learning algorithm has many parameters ([FI05]; [RdC08]). Thus, parameter optimization is often accomplished using a Grid Search on discrete sets of parameters to select the optimal ones with the aid of cross-validation [FA19]. This method basically test all parameters combination and measuring the efficiency according to the metric chosen by the developer. In the current research, accuracy was chosen as the metric to measure the performance of the combination of all hyperparameters, thus, the hyperparameters combination that returned the highest value of accuracy was used to set up the models. The hyperparemeters that were utilized to feed the Grid Search were, $C$ (the penalty of the error), and $tol$ (it is a stopping criterion), both hyperparameters with input values of, 0.001 - 0.01 - 0.1 - 1.0 - 10 - 100 - 1000. The optimal hyperparameters values returned by the Grid Search were $C = 100$ and $tol = 0.001$.

#### 3.1.2.2 Neural Networks

Neural networks are a field of Artificial intelligence (AI) in which mathematician algorithms have been worked in a process that remember the brain thinking process. Neural network models are biologically plausible and can help us understand how the brain works [MMK03]. The elements that compounds a Neural Network are illustrated in the figure 3.2. The figure 3.2 show us a biological neuron on the left and an artificial neuron on the right.

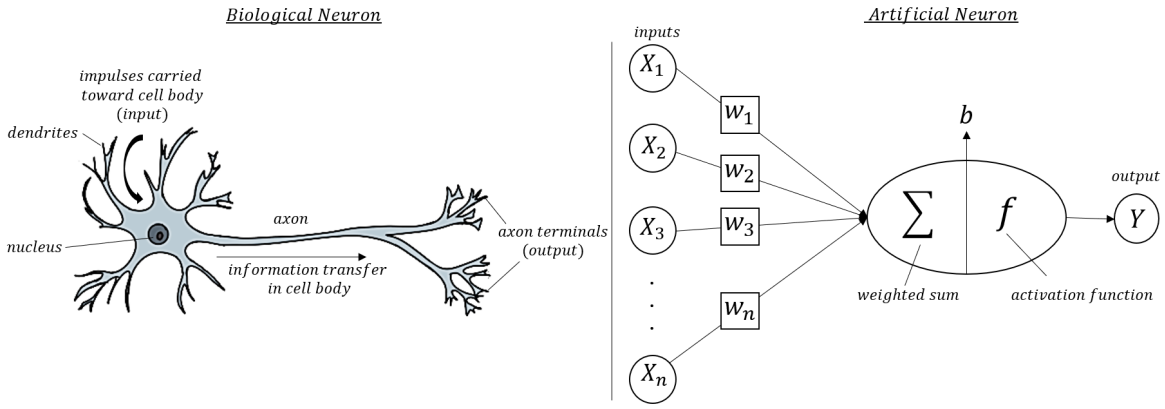



Figure 3.2: Biological and artificial neurons showing the similar compounds.

There are some points to pay attention:

(i) The first point are the input nodes, is through them that the model is feed with the input data, in recognize problems this nodes can be understood as the object features

(ii) Each connection between the input nodes and the hidden layer nodes has a weight associated with connection, based on the individual weight of each connection the algorithm selects which input features will have greater or lesser impact on processing.

(iii) The weighted sum is a calculus that can be represented by the equation 3.1.7, to calculated the weighted sum is necessary, the connection weight and the input node value. The weighted sum is expressed by the equation below:

$$\textbf{Weighted Sum} = b + \sum_{i=1}^{n} X_i * W_i \tag{3.1.7}$$

Where $X_i$ is the input node value, $W_i$ the weight associated to connection and $b$, bias, an additional weight that allows to move the activation function to the left or right to improve model learning.

(iv) The result obtained with the weighted sum is transfer to activation function. The activation function is chosen according to model goal.

(v) The output node is the last compound of the default Artificial Neuron, it is function of the all elements listed before and can be expressed as follow in equation 3.1.8:

$$\textbf{Output node} = Y = f(b + \sum_{i=1}^{n} X_i * W_i) \tag{3.1.8}$$

The process describe above can be understood as feed-forward, which is the information flow in the neural network from input data to the output data [SKP97]. The feed-forward doesn't allow the update of the weights, because this is necessary the application of the back-propagation. The discovery of "back-propagation" in the context of neural networks by Rumelhart et al. [RHW86] drastically improves the learning efficiency of such models enabling them to be used in practice [Ba13].The central idea of error back-propagation is to compute the partial derivatives of the weights in a neural network by applying the chain-rule repeatedly [Ba13].

The learning rate can be understood as the amount that is updated in each epoch in the weights of the model. It is used to find the best combination of weights so that the model reaches the minimum error in its predictions. The procedure of choosing a good value for the learning rate of the stochastic optimization can increase performance and reduce the training time in the model. This because if were chosen a value for the learning rate bigger than expected, the model can update the weight more than necessary and occasionally jump the minimum value of error, in this case, the model will have a needless time consuming. On the other hand, however, if the value for the learning rate is less than expected, the model will have to update several times until it reaches the minimum error value, and again this process will be unnecessarily time-consuming. The learning rate can be considered the most important hyperparameter, thus, choosing a good value for the learning rate is fundamental in training deep learning neural networks.

The main purpose of the activation function it is to adds the nonlinear factors to remove redundant data while preserving features, it retains"active neuron feature" and maps out these features by nonlinear functions, which is the essential of the neural network to solve the complex nonlinear problem [LS18]. The Relu (Rectified Linear Units) activation function [NH10], [HSM+00], [BDL20]is a mathematical function whose output 0 if the input is negative, and for any positive value x, it returns that value back. In neural networks modeling of electrofacies, predicting the probability of the electrofacies requires computing scores for every electrofacies in the data set and to normalize them to form a probability distribution. This is typically achieved by applying a softmax function. The Softmax activation function [WLZ+18], [JCGJ17], is often used in the output layer of neural network models for multi-class classification problems, where the number of output classes is required on more than two class labels. Softmax makes this possible because

the output is a vector with the probabilities of occurrence of each class label.

The hyperparameter, batch size, used in the model corresponds to the number of training samples that will be executed before the weights update. The advantages of using a batch size smaller than the samples number is are two.

(A) The first advantage is because it required less memory, when the data set is very large this is a crucial factor.

(B) Normally, Neural Network train is faster, in addition, neural network parameters will have more than just an update because the model will have updated neural network parameters after each batch size. In the case that, the batch size is equal to the sample number, the model will only have an update of the Neural Network parameters.

Several recently proposed stochastic optimization methods have been successfully used in training deep networks such as RMSPROP, ADAM, ADADELTA, NADAM. They are based on using gradient updates scaled by square roots of exponential moving averages of squared past gradients [RKK19]. In the model applied in the research, the optimizer chosen was the Adam optimizer due to its lower training cost and faster convergence in relation to the other optimizers. The name Adam derived from adaptive moment estimation. Adam is a method for efficient stochastic optimization that only requires first-order gradients with little memory. The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients [KB14]. Therefore, Adam is an optimizer that can be used as a good alternative to the classic Stochastic gradient descent, the Adam optimizer is called when upon updating neural network weights.

In spite of many successes, Neural Networks still suffer from a major weakness. The presence of nonlinear hidden layers makes deep networks very expressive models which are therefore prone to severe overfitting [Sri13]. Due to this reason, was applied the dropout technique to the Neural Networks model used in the research. The dropout is a technique used in neural networks to try to avoid the overfitting of the algorithm. The key idea is to randomly drop units (along with their connections) from the neural network during training [SHK+14]. Besides to avoid the overfitting, the technique called dropout has shown to significantly improve the performance of deep neural networks on various tasks [HSK+12].

For running the model developed in the research was utilized the Google Colab with

the TensorFlow 2.0, version 2.5.0, the main library used to build the model was the Keras, version 1.1.2.

## 3.2 Geophysical Well Logs

Geophysical well logging consists, for the most part, of lowering instrument packages into holes in the earth in order to measure physical parameters that characterize the formations [Tit12]. Measurements techniques are used from three broad disciplines: electrical, nuclear, and acoustic. Usually a measurement is sensitive either to the properties of the rock or to the pore-filling fluid [ES07]. In current research, the geophysical well logs used were, Gamma ray, Neutron, Resistivity, Water saturation, Density, Porosity, P and S-velocity. They have contributed with the construction of the data set, and were utilized as input features for the building of models of the research.

The geophysical well logs were chosen according to its capacity of representation of the rock matrix, porosity and fluids. This choice was made with the goal of understanding all elements presents in the rock interval of interest. Between the groups of geophysical well logs, able to represent the rock matrix, porosity and fluids, were chosen the geophysical well logs in common between the well logs worked in the research. In the section 4.1 it is explained more clearly the reasons why were chosen the current setup of geophysical well logs.

### 3.2.1 Gamma Ray Logging

Gamma Ray (GR) logs were introduced in mining research around 1930, and, ever since, have been utilized to auxiliary the interpretation and correlation of lithologies and formations. Gamma Ray (GR) logs measure the natural radioactivity in Formations. In addition to prospecting for radioactive minerals, the radioactive method is extensively applied in borehole studies of subsurface stratigraphy as might be deemed necessary when prospecting for crude oil and natural gas [RWS+14]. Shale-free sandstones and carbonates have low concentrations of radioactive material and give low Gamma ray readings. As shale content increases, the Gamma ray log response increases because of the concentration of radioactive material in shale [AKG04]. Shales usually emit more Gamma rays than other sedimentary rocks because radioactive potassium is an abundant component in their

clay content, furthermore, due to the cation properties of clay, uranium and thorium are absorb in its structure.

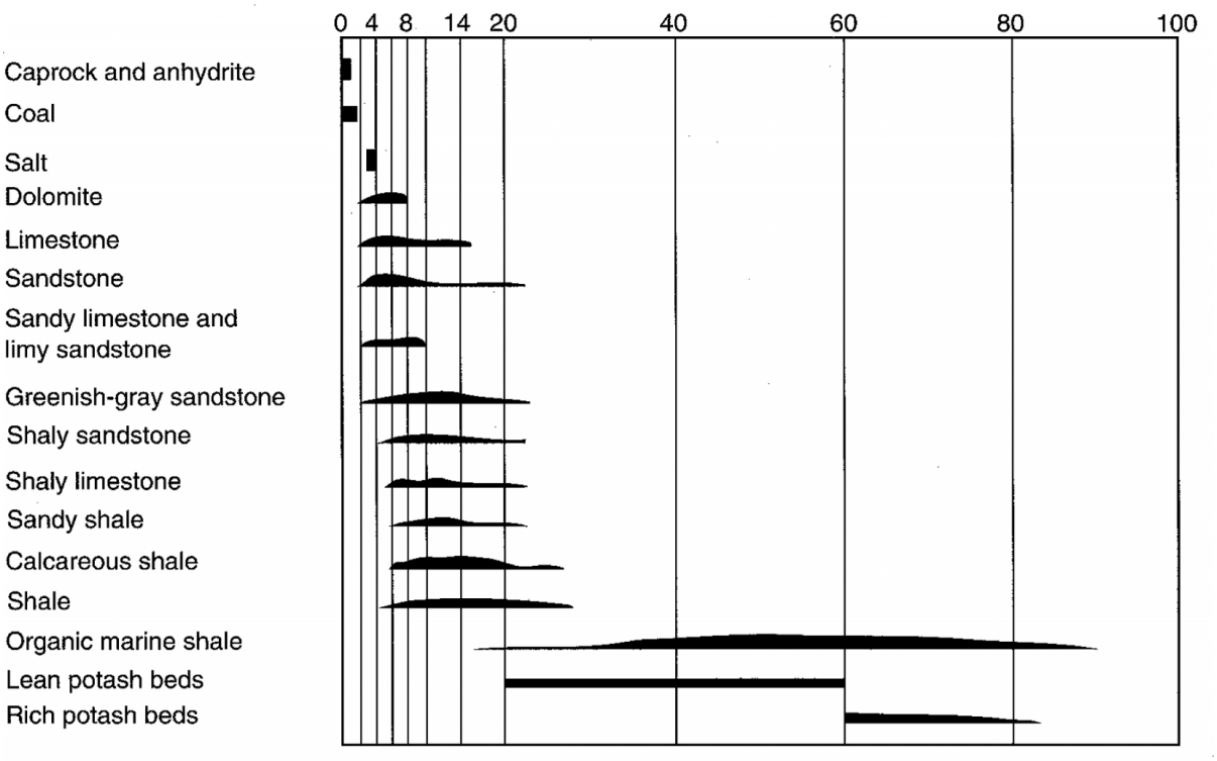



Figure 3.3: Gamma Ray distribution in the rocks (from [Big92] after[Rus44].

In the figure 3.3, it is possible to see the distribution of Gamma Ray in rocks. Gamma Ray logs have mainly used to distinguish clean and dirty formations and, consequently, potentially productive intervals to hydrocarbons. Therefore, it is a good indicator of reservoir zones and non-reservoir zones.

Gamma ray logs has the mainly use to distinguish clean and dirty formations and, consequently, potentially productive intervals to hydrocarbons. Therefore, it is a good indicator of reservoir zones and non reservoir zones.

### 3.2.2 Neutron Logging

In Neutron logging, a neutron source is employed to excite the release of radiation from the rock formations. The neutron source is usually a mixture of elements (of which of beryllium and radium have been commonly used) and the method is a means for determining the relative porosity of rock formations and to auxiliary in the detection

of light hydrocarbons and gas [RWS+14]. The Neutron device works by bombarding the formation with high-energy neutrons. When these high-energy neutrons collide in formations with a large amount of hydrogen (atoms with similar size to neutrons atoms), the neutrons can interact in two ways, the first is in the collision with hydrogen atoms, and the second is in the absorption of neutrons by the nuclei of atoms of the formation. In both neutrons interactions, there is the production of high-energy Gamma rays, and in the collision, there are low energy neutrons being scattering. Due to the higher concentration of hydrogen in the fluids present in the pores of the rock formation, the neutron logging response indicates the concentration of hydrogen in the fluid-filled pore space. Thus, the neutron log can be used to determine the hydrogen content of the logged interval by counting captured gamma rays or neutrons counted at a detector [Fan10].

It is known that Water and Oil have high concentrations of hydrogen when compare with Rock formations and Gas. Therefore, it is possible to infer that the neutron detectors will have a large Gamma ray response in formations with high porosity fluid-filled in Oil and Water, while formations with a small Gamma ray response will have a low porosity or the porosity fluid-filled with Gas. The Neutron logs are widely utilized, but the main uses are in qualitative identification of shale/sands zones, in cross plots with Density logging to detect gas zones, light hydrocarbon zones, and to determine lithology.

### 3.2.3 Density Logging

Density log tools consist of a chemical radioactive source of Gamma ray, Cesium and Cobalt, it is responsible by the Gamma ray emission of high energy. The interaction and the number of collisions with electrons of the rock formation is proportional to the electrons number of the rock formation, thus, the electrical density can be related with the volumetrical density. If the measurement of Gamma rays is large, this implies a large electron density and correspondingly large bulk density [FS16]. If more Gamma rays were absorbed in the rock formation, the ray count transmitted to the detector will be less and, therefore, the rock formation will be less dense. The Gamma ray that collide with the materia can suffer three types of interaction: (i) production of positron-electron pars; (ii) Compton scattering; (iii) and Photoelectric effect. The Compton scattering is the main effect used to measurement the bulk density.

The electrical density is frequently calculated by the equation (3.2.1) as follow:

$$\rho_e = 2\rho_b \frac{Z}{A} \tag{3.2.1}$$

Where $\rho_e$ is the electrical density, $\rho_b$ is the bulk density, $Z$ is the atomic number, and $A$ is the molecular weight. The formation porosity of a interval can be calculated by equation (3.2.2) below:

$$\phi = \frac{\rho_{ma} - \rho_b}{\rho_{ma} + \rho_f} \tag{3.2.2}$$

Where $\phi$ is the porosity of the interval, $\rho_{ma}$ is the matrix density, and $\rho_f$ the fluid density.

The traditional use of Density logs is to measure the density and porosity of reservoirs [SJ93]. Density logs are useful for determining hydrocarbon density and in detecting hydrocarbon gas when associated with Neutron logs, a small density implies high hydrocarbon gas content, while a high density suggests a low hydrocarbon gas content [FS16]. Density logs are also helpful for lithologic identification, creation of Synthetic Seismograms when combined with Sonic logs and mensuration of Formation mechanical properties when associated with Sonic logs. In the literature is suggested running a caliper log to complement interpretation of the density log. Because the density is sensitive to borehole rugosity, a correction curve should always be run [Flo13].

### 3.2.4 Resistivity Logging

The Resistivity logging tool was first used downhole in Alsace, France in the year of 1927 [SI16]. The Resistivity is the ability of a material resists to the flow of electrical current. The ability to conduct electric current depends of the rock matrix and of the fluids contend in the pores. Usually the rock matrix presents difficulties into the flow electrical current. The pores when non-filled working as isolating. When there are free ions in the fluid content of the pores as occurs in a brine, then the flow electrical current is easier. The usefulness of electrical resistivity logging rests on the fact that rocks (with a few notable exceptions) and hydrocarbons are insulators, whereas connate waters are generally saline and, therefore, good conductors [Tit12]. In the geological prospect of an Oil field, the main interest lies in the fluid that the formation can produce. Therefore, it is crucial to

know distinguishing between brine and hydrocarbons, the two main fluids that occupy the pores space [ES07]. Resistivity differences between brine and hydrocarbons make it possible to use resistivity logs to distinguish between brine and hydrocarbon fluids [FS16]. Thus, high values of resistivity allying with low values of Water Saturation and high values of porosity, can indicate rich zones in hydrocarbons (reservoir zones), while low values of resistivity allying with high values of Water Saturation and high values of porosity implies the presence of brine, and consequently non-rich zones in hydrocarbons (non-reservoir zones). The Water Saturation can be calculate from the Resistivity logging from the equation (3.2.3) The term "Water Saturation" is frequently used to describe the fraction of the rock pore space that is filled with water. The Water Saturation also can be defined as the percentage of the porosity occupied by brine rather than hydrocarbons[ES07].

The Water Saturation ($S_w$) can be calculated from Archie's equation (3.2.3) for wetting phase saturation,

$$S_w = \left(\frac{F \cdot R_w}{R_t}\right)^{\frac{1}{n}} \tag{3.2.3}$$

where $F$ is the formation resistivity factor, $R_w$ is the resistivity of the ionic solution, $R_t$ is the formation resistivity, and $n$ is called the saturation exponent. The Water Saturation when allied with the Resistivity logging becomes a powerful guideline in Oil and Gas prospecting.

### 3.2.5 Sonic Logging

Sonic logs is another technique used to measure porosity, working with measures that reflect the elastics characteristics of the rock formation, the technique is widely utilized in literature and industry. The attribute measurement, interval transit time, is relative to the elastic properties of the rock formation, such as, lithology (matrix composition), texture, fluid contend and porosity. In other words, Sonic logging is a measurement of mechanical wave slowness (the inverse of the velocity) throughout the formation, produced by a source located inside a tool immersed in a fluid filled borehole [SBNR19]. However, it is important to highlight that the quality of the Sonic logging is highly dependent on borehole conditions, and requires good contact between the tool and the borehole wall [SR11]. The main functions of the Sonic logs are, evaluating the rock formation porosity,

lithological identification when associated with Density or Neutron logs, construction of Synthetic Seismograms when allied with Density logs, and permeability identification.

The ratio of compressional-wave velocity ($V_p$) to shear-wave velocity ($V_s$), or $V_p/V_s$ gives additional information about lithology. Well logs studies ([Pic63]; [Nat74]; [Kit77]; [MS90]) indicates a correlation between $V_p/V_s$ values and lithology. Pickett [Pic63] established $V_p/V_s$ values from core measurements of 1.9 for limestone, 1.8 for dolomite, and 1.6 to 1.75 for clean to calcareous sandstones [PF97]. The acoustic logs also can help in the distinguishing of the consolidated, unconsolidated, homogeneous and fractured formations, it is important because this mechanical property will influence in the stability of the borehole and in the permeability of the formation. From Whylle's equation (3.2.4) [WGG56],

$$\phi_s = \frac{\Delta t_{log} - \Delta t_{ma}}{\Delta t_f - \Delta t_{ma}} \tag{3.2.4}$$

where the terms are: $\phi_s$, the porosity of the formation; $\Delta t_{log}$, the acoustic interval transit time in the formation of interest; $\Delta t_{ma}$, the acoustic transit time of the rock matrix and $\Delta t_f$, the acoustic transit time of interstitial fluids. It is possible to note that, the porosity calculated from Sonic logs will depend on the choice of the acoustic transit time of the rock matrix, which varies with lithology. For unconsolidated and poorly consolidated formations a correction factor is required in order to adjust the Whylle's equation (3.2.4). Thus, rewriting the Whylle's equation (3.2.5):

$$\phi_s = \left(\frac{\Delta t_{log} - \Delta t_{ma}}{\Delta t_f - \Delta t_{ma}}\right) \times \frac{1}{C_p} \tag{3.2.5}$$

Where $C_p$ is the compaction factor, the term added in the Whylle's equation (3.2.4) for unconsolidated formations.

Another equation used to calculate the porosity from Sonic logging was developed in Raymer and Hunt works [RHG+80], the equation is known as Raymer-Hunt equation (3.2.6), expressed as,

$$\frac{1}{\Delta t_{log}} = \frac{\phi_s}{\Delta t_f} + \frac{(1-\phi_s)^2}{\Delta t_{ma}} \tag{3.2.6}$$

This provides a much superior accuracy porosity over the entire range of geologically reasonable $\Delta t_{log}$[Glo00]. From Sonic logging is possible to note that sound waves propagate faster in rock matrix than in fluid. Thus, Rock Formation's that have long interval transit time imply slow speed of sound propagation and large pore space. Conversely, short transit time implies a high speed of sound propagation and small pore space [FS16]. Therefore, the Sonic logging is a useful tool to calculate porosity and for auxiliary others logs in the identification of the Rock Formation's physical properties in the subsurface.

# Chapter 4

# Materials and Methods

The chapter will show the materials and methods used in the research. To realize the predictions were utilized the softwares, Jupyter notebook, an open-source software with Python version 3.8.3, and Google Colab with TensorFlow 2.0 version 2.5.0. The software used to extract the input features and to plot the Geophysiscal well logs section was the Petrel, version 2017. The methods applied in the research can be visualized in the next sections. The section 4.1 talk about the data set, all the issues and solutions found by the way, how the data set was acquired, and the process of features selection. The section 4.2 and its subsections are responsible to presents the data science stage, in this stage the depth matching process 4.2.1, and the Statistical processing 4.2.2 are described. The splitting stage can be visualized in the section 4.3, this section presents how the data set was splitting into training and test data set. In the section 4.4 are described the workflow of the research in the cases of supervised and unsupervised learning algorithms. In the current research was proposed two different way to make the electrofacies classification, the first (i) workflow was thought to supervised learning algorithms (section 4.4.2), while the second (ii) was developed for unsupervised learning algorithms (section 4.4.3). In the section 4.5 it is described all the process of statistical validation that was used in the supervised learning algorithms to evaluate the models. The last section (4.6) of this chapter will describe the vertical heterogeneity, an metric created in the research that measure the purity level of reservoir zones.

## 4.1 Data set

The research worked with data of Geophysical well logs presents in a Brazilian pre-salt oil field, Gato do Mato. The two wells worked in the research were, 1-SHEL-23-RJS and 1-SHEL-26-RJS, the data set of the both wells were utilized in the unsupervised and supervised learning algorithms. The data set of Geophysical well logs was provided by the Brazilian National Agency of Petroleum (ANP). After the data acquisition stage, the next stage was the features selection. How the research is working with the prediction and classification of electrofacies, the feature selection stage can be understood as the choosing of the best Geophysical well logs capable to distinguish between the different electrofacies, and at same time being able to represent the characteristics of the rock matrix, fluids and porosity. There are several Geophysical well logs capable of measuring these three properties (rock matrix, porosity, and fluids), but to avoid problems of Geophysical well logs absence in the data set, the research picked out Geophysical well logs in common between the wells worked in the current research. The Geophysical well logs used were, about the rock matrix: Gamma-ray, and S-velocity. About the porosity: Neutron, Sphi, and Density. About fluids: Resistivity, and Water Saturation. These are the seven input features responsible for informing about the properties of the rock formation. Initially, the P-velocity input feature was chosen, but as will be shown in the section 4.2.2, this input feature was removed. Once that Geophysical well logs were selected, the data science stage 4.2 has started. At this stage was made the depth matching process 4.2.1, and the statistical processing 4.2.2 in the data set. Both stages are illustrated in the next sections.

## 4.2 Data science

As a core of the research, the data science stage was concentrated in the creation and treatment of the data set. In the current research the term data science was used to describe the process of creation and treatment of the data set. The building of the data set is an essential stage and must be done with expertise because it is the guarantee of reliable information. In the data science stage was realized the depth matching process, and statistical processing in the data set.

To carry out the data science stage, the research used Jupyter notebook, Google Colab, and Excel as auxiliaries tools in the building process of the data set. They were

responsible for ensuring efficiency, agility, and precision in manager with the data set.

### 4.2.1 Depth matching process

The challenge in the depth matching process is to match the seven input features in-depth. Each input feature is provided individually by the ANP, and many of these input features (geophysical well logs) do not have records at the same depth. Thus, creating an algorithm capable of building a data set with the seven matching records in depth it is a crucial step for the next stages of the project. The depth-matching process of the seven input features in relation to depth was responsible for a significant reduction in the number of input records. The Gamma ray and Density logs demonstrates very well this reduction, initially the number of Gamma ray collected was 7277, and the number of Density records was 11707. After the application of the depth-matching algorithm in the seven input features, the number of records decreased to 1624 input records. The logging data sampling interval of the data set is 0.15 meters. Therefore, by multiplying the number of records by the data sampling interval, it is possible to find 243 meters approximately of rock formation in the 1-SHEL-23-RJS well. This rock formation interval was used to work on the training and testing of the models presents in the current research.

### 4.2.2 Statistical processing

After the depth matching process, the data set was normalized in order to transform all geophysical well logs in the interval between 0 and 1. This is an important process because it prevents that the absolute values of the records have a weight higher or smaller. In order words, the normalization process makes the model training less sensitive to the scale of input features. Initially were chosen eight input features, these input features were normalized and their frequency histograms were plotted as can be seen in the figure 4.1.

The procedure for the construction of frequency histograms was made due to the outlier detection method, when the data distribution has a normal distribution behavior it is common to use the method mean plus or minus two/three standard deviations [Mil91]. In this method the values that are outside of this range are considered outliers. When it is used the mean plus or minus two standard deviations, 95.45% of the data are present in the range. The method of the mean plus or minus three standard deviations is based on the characteristics of a normal distribution for which 99.87% of the data appear within

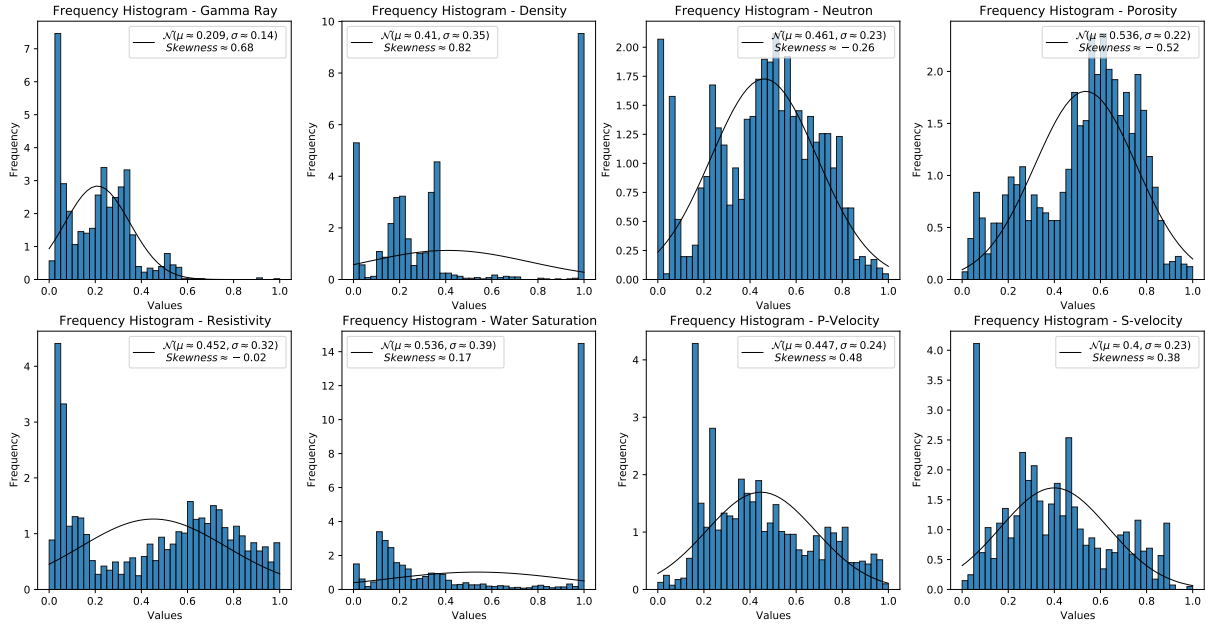



Figure 4.1: Frequency histogram of the Geophysical well logs. Data set normalized.

this range [HRYB98]. But, it is important to remember, these results just are valid in the normal distribution case. Beyond of the visual aspect of the frequency histograms, the research also calculated the Fisher-Pearson coefficient of skewness to auxiliary in the analysis. In the figure 4.1, frequency histograms are plotted with the calculus of the mean($\mu$), standard deviation($\sigma$), and the Fisher-Pearson coefficient of skewness.

The sample skewness is computed as the Fisher-Pearson coefficient of skewness [KZ00]. The Fisher-Pearson coefficient of skewness is responsible to measure the lack of symmetry in a distribution. Normal distribution has skewness 0. Larger values (in magnitude) indicate more skewness in the distribution of observations [KZ00]. Thus, the Fisher-Pearson coefficient of skewness it is useful to help in detection of normal distribution. The Fisher-Pearson coefficient of skewness was applied in the eight input features, and the values can be seen in the figure 4.1. It is possible to note from figure 4.1 that the population of the input features have unknown distributions, and with a notable asymmetry (skewness). Thus, the method of outliers detection based in the mean plus or minus two/three standard deviations is not valid to use in the research data set. Due to the notable asymmetry and the unknown distributions illustrated in the figure 4.1 was necessary to find another method capable of detecting the outliers with reliability.

One first approach would be to make the data symmetrical by the use of a non-linear transformation, in the research case, the logarithmic transformation was used. Presum-

ably with such an approach, outliers would also be more symmetrically away from the central tendency, providing an equal chance of locating low and high outliers. However, there is no single technique that makes the data symmetrical when they have been contaminated with outliers [CC10]. How can be seen in the figure 4.2, the logarithmic transformation does not have successful in to become the distributions normal. But, in general, the values of skewness of the frequency histograms have decreased. The exceptions were the Resistivity, Water Saturation, and P-Velocity distributions. These input features had the absolute magnitude of skewness values increased.

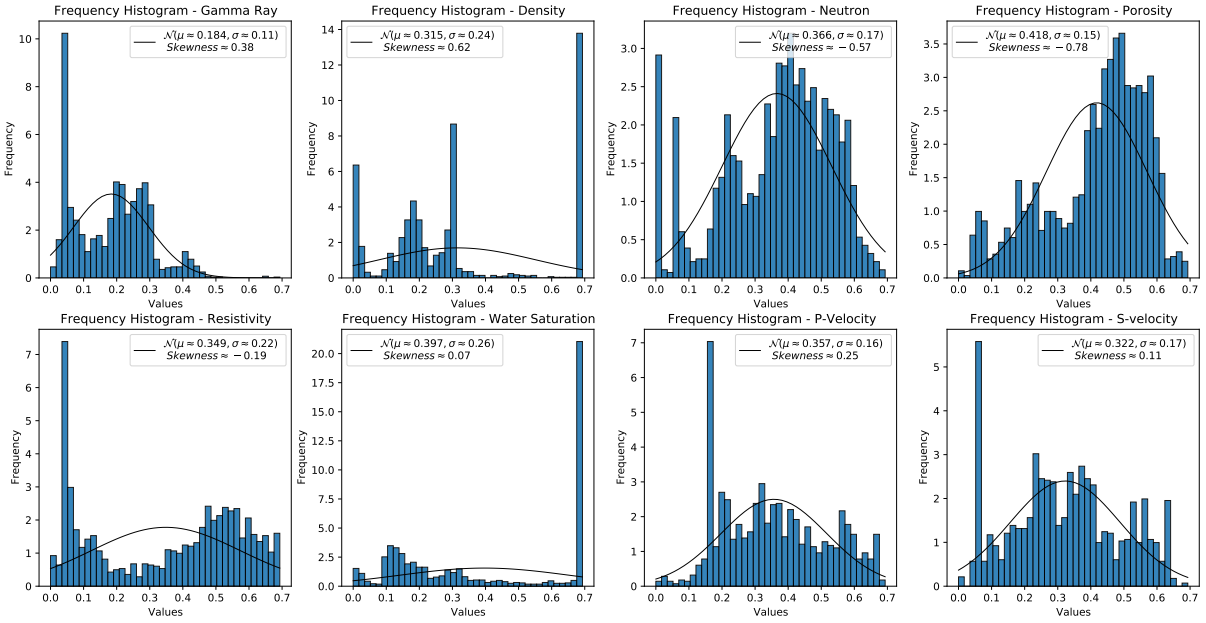



Figure 4.2: Frequency histogram of the Geophysical well logs. Data set normalized and log-transformed.

Once that, the logarithmic transformation was not capable to transform the distributions as normal distributions, the method Inter-Quartile Range (IQR) method of outlier detection was applied [Daw11], [LJK+00], [VPS18]. The method Inter-Quartile Range (IQR) method of outlier detection is usually used when the distributions are not normal and asymmetrical. The Inter-Quartile Range (IQR) is the difference between third and first quartile. Quartile are responsible to divide the ordered sample observations into four quarters having the same number of observations (m) in each quarter [JF01].

The skeletal boxplot consists of a box extending from the first quartile ($Q_1$) to the third quartile ($Q_3$); a mark at the median; and whiskers extending from the first quartile to the minimum $Q_1-(1.5\cdot IQR)$, and from the third quartile to the maximum $Q_3+(1.5\cdot IQR)$

[Daw11].

The bloxplots of the distributions were created in Jupyter notebook and Google colab 4.3. In the method Inter-Quartile Range (IQR), method of outlier detection, the samples that are outside of the interval between the minimum $(Q_1 - (1.5 \cdot IQR))$ and the maximum $(Q_3 + (1.5 \cdot IQR))$ are considered outliers.

It is possible to note in the figure 4.3, that boxplots of Gamma ray and Porosity distributions has presented outliers respectively above the maximum value, and below the minimum value.

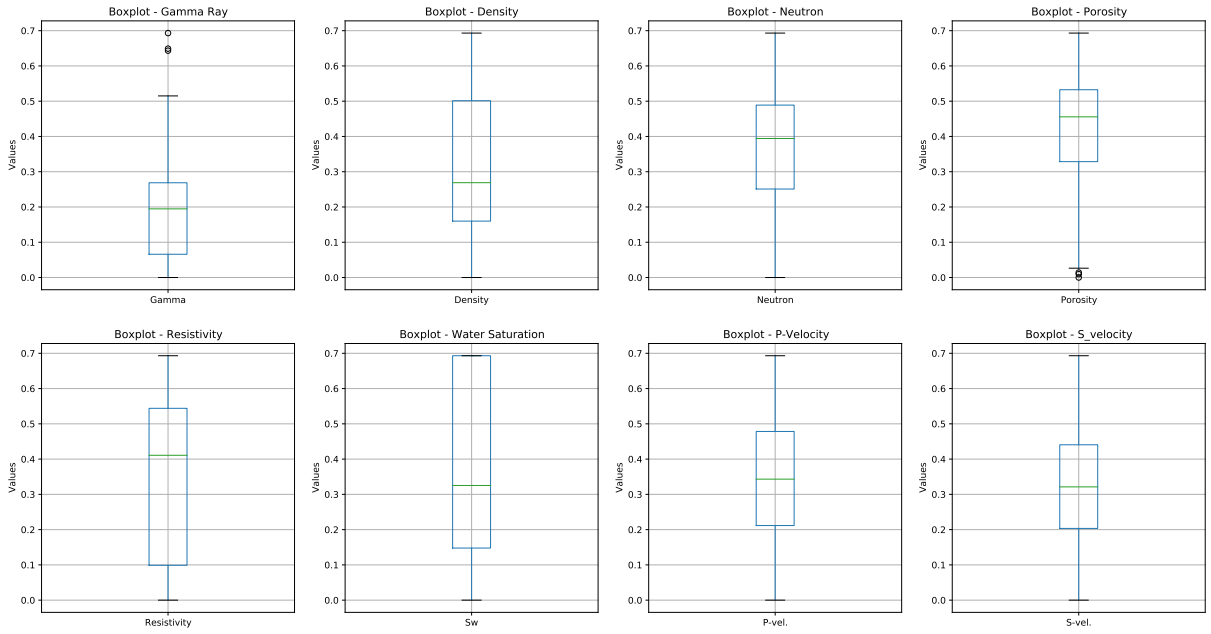



Figure 4.3: Boxplot of the Geophysical well logs. Data set normalized and log-transformed.

Once the outliers were detected, an algorithm created in the research was executed to remove the rows from the data set that contains the outliers. Before to remove the outliers, there were 1624 records in the data set, and after the outliers removal the numbers of records in the data set decreased to 1617 records. Among the removed data, three were porosity outliers, and four were gamma ray outliers. The porosity outliers were detect below to the minimum value in the boxplot, while the gamma ray outliers were detect above to the maximum value in the bloxplot. The final data set without the outliers can be seen in the figure 4.4.

After detecting the outliers, there is only one more step to be made to become the algorithm more optimized. The last step it is to verify if the input features are independent

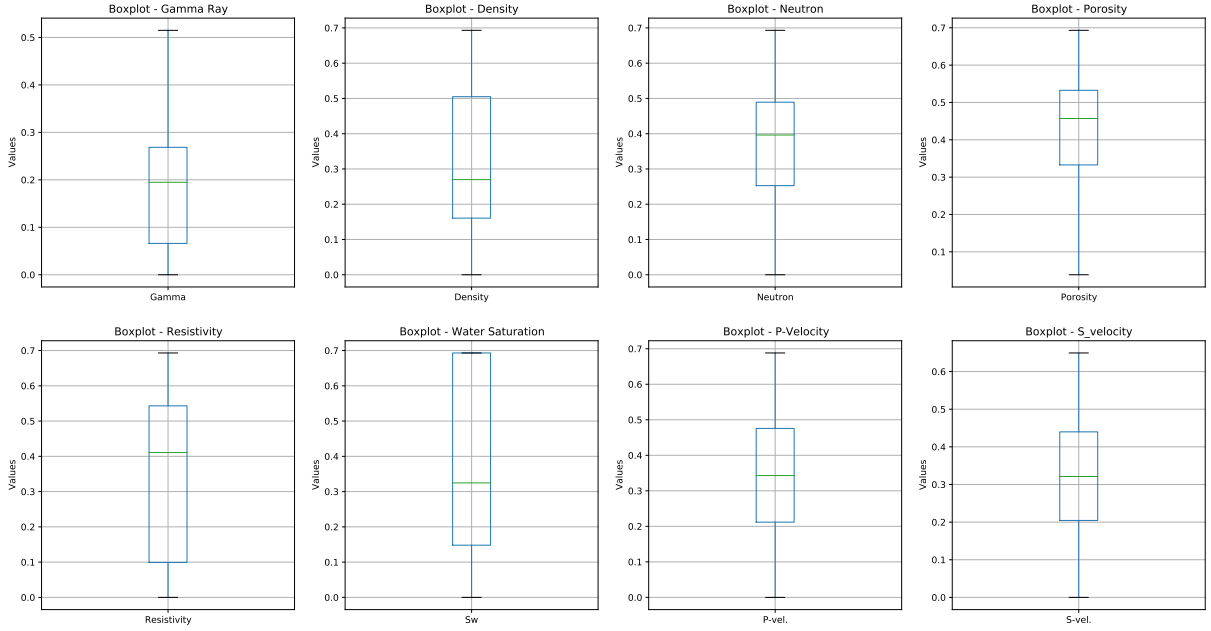



Figure 4.4: Boxplot of the Geophysical well logs. Data set normalized and log-transformed and outliers removed.

between them self. This important step avoids that the model working with redundant data (dependent variables). In this test it is calculated the Pearson correlation coefficient [BCHC09]. If the Pearson correlation coefficient is 1, the variables are totally dependent on each other. If the Pearson correlation coefficient is -1, the variables will have a total inverse dependence. In the other side, if the Pearson correlation coefficient is equal 0, the variables are independent or uncorrelated. In the research, the Pearson correlation coefficient of all variables was calculated, then with the values found a matrix was built, this matrix with all values plotted is known as correlation matrix. The correlation matrix of the variables (input features or geophysical well logs) can be seen in the figure 4.5.

From the figure 4.5 it is possible to note that the correlation matrix between the geophysical well logs presents a linear dependency between the two variables P-velocity and S-velocity. The value of the Pearson correlation coefficient for these two variables it is equal 1. Thus, for optimizing the algorithm one of these variables needs to be removed. In the research case was removed the variable P-velocity. The correlation matrix after the removal can be seen in the figure 4.6

Therefore, after the statistical processing stage (last stage in "Data science stage"), the data set it is finalized and ready to be used in models.

One difficulty with treatments of outliers is that there is no unanimously accepted

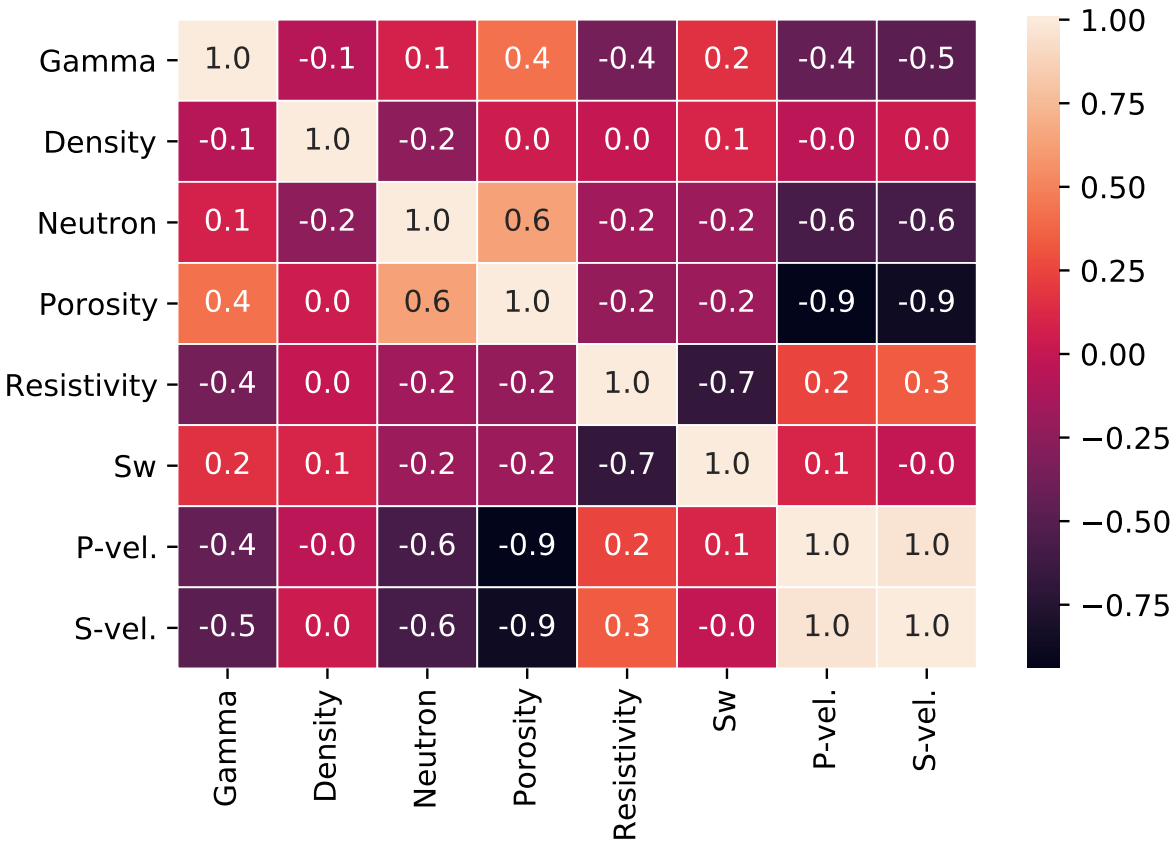



Figure 4.5: Matrix Correlation between the Geophysical well logs. Data set normalized and log-transformed and outliers removed.

theoretical framework for the treatment of outliers. Various fields have developed various approaches and rare are the approaches that can be formulated with the concepts of another approach [CC10]. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation, and incorrect results. Therefore, it is important to identify them prior to modeling and analysis [BG05], [LSJ04], [WBH+02]. In the research, the Inter-Quartile Range (IQR) method was chosen, which showed, through boxplot analysis , a good adaptation for detecting outliers in the data set.

It is important to remember that, the initial efforts to work with the data set at the data science stage, are the core of researches related to the application of AI algorithms. When the building, and treatment of the data set is well-realized, AI algorithms application become more efficient, in some cases can increase the hit rate and consequently lead to better results. The evolution of the research in models I and II just was possible to achieve due to the precise development realized in the data science stage.

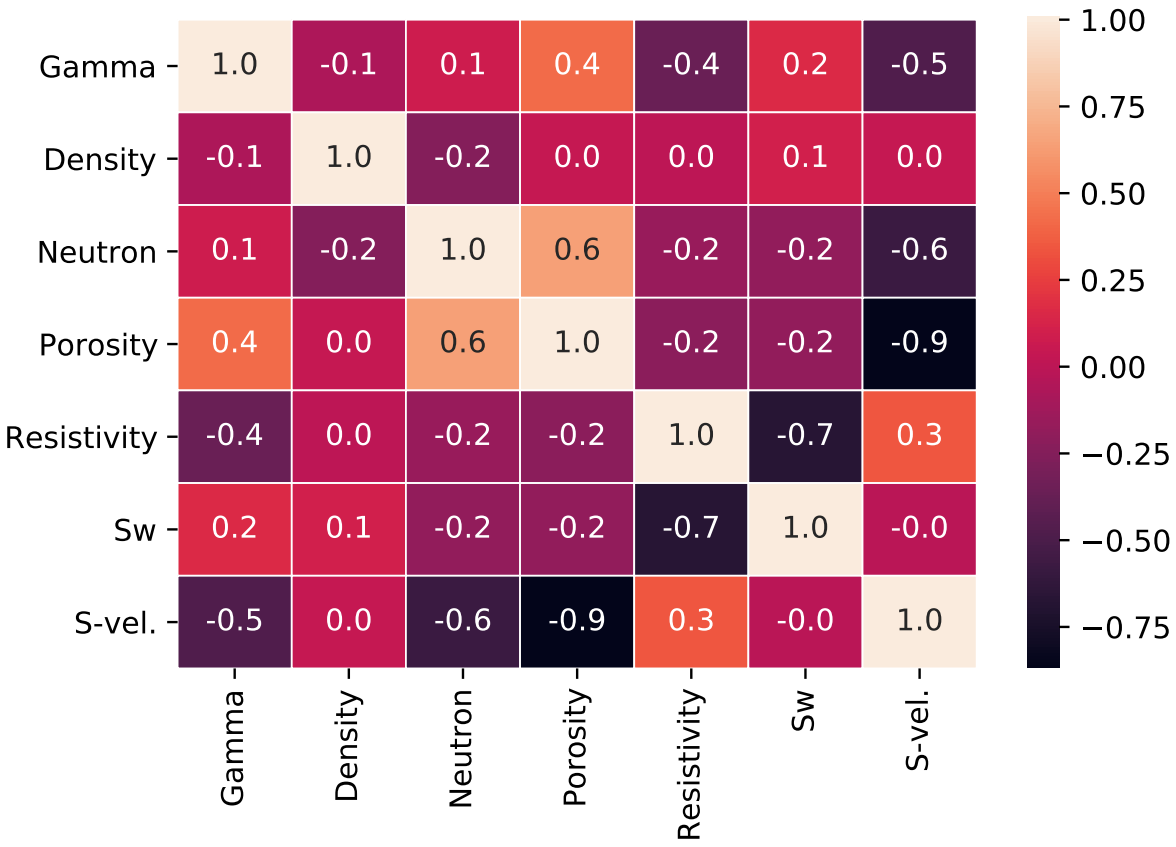



Figure 4.6: Matrix Correlation between the Geophysical well logs without P-velocity. Data set normalized and log-transformed and outliers removed.

## 4.3 Training and Test data set

The main data set, the 1-SHEL-23-RJS data set, was created according to a manual classification made by a specialist Geologist, this classification was responsible for labeling the data set. The labeling process is necessary because it allows to evaluate the performance of the models. The data set consists of 1617 inputs, and each input has 7 distinct features, they are: Gamma-ray, S-velocity, Neutron, Sphi, Density, Resistivity, and Water Saturation. All these features, Geophysical well logs, are responsible by recorded an information about the Rock matrix, fluids and porosity of rock formations in subsurface.

With 1617 records in the final data set, and a logging data sampling interval of 0.15 meters it is possible to calculate the interval of rock formation worked in the research. Multiplying the number of records by the data sampling interval, approximately 243 meters of rock formation are found in the 1-SHEL-23-RJS well. This rock formation interval was used to work on the training and testing of the models presents in the current

research.

In models developed with Supervised Learning algorithms, a way to evaluate the model performance it is dividing the labeled data set into two groups, the training, and test data set. In the current research, this division was made only for models built with Supervised Learning algorithms. In models developed with Unsupervised Learning algorithms, there is no meaning in dividing the data set, once that the data set is unlabeled.

In the current research, the training and test data set were divided respectively in the proportion of 75% and 25%, in the 1-SHEL-23-RJS data set, the division was made at random and proportional to each electrofacies labeled in the manual classification stage (Table 4.1).

Table 4.1: Number of samples of each electrofacies selected to train models I and II.

| Electrofacies | Proportion of Samples |
|---|---|
| Electrofacies 0 | (125/507) 0.246 |
| Electrofacies 1 | (025/087) 0.287 |
| Electrofacies 2 | (125/532) 0.235 |
| Electrofacies 3 | (087/333) 0.261 |
| Electrofacies 4 | (044/165) 0.267 |

The random division is necessary because it prevents that the model selects only a part of the data set, which can generate an addiction to the model in a specific data set interval. The training and test data set is present only in the 1-SHEL-23-RJS data set. The training data was used to teach the models, while the test data was used to make the statistical validation. The 1-SHEL-26-RJS data set was used to prediction of the electrofacies learned in the 1-SHEL-23-RJS training data set.

## 4.4 Working Flowchart

The proposed working flowcharts were thought with the goal of becoming the electrofacies classification an automated process, and then, save time in the process of electrofacies classification. To achieve the main goal, the research proposed the application of Artificial Intelligence (AI) algorithms to make the classification and prediction of electro-

facies in carbonate rocks. This process usually gets to achieves a high hit rate and ensures a standardized in the electrofacies classification and prediction. In the current research was worked supervised and unsupervised learning algorithms, and among them the working flowcharts for each one is different. In the case of supervised learning algorithms the research proposed initially a manual classification in only one well (1-SHEL-23-RJS), this well was called in the research of well reference (Figure 4.7), it was utilized as a model to teach the AI algorithms how to make the classification and predictions in new wells inside of the same petroleum field.

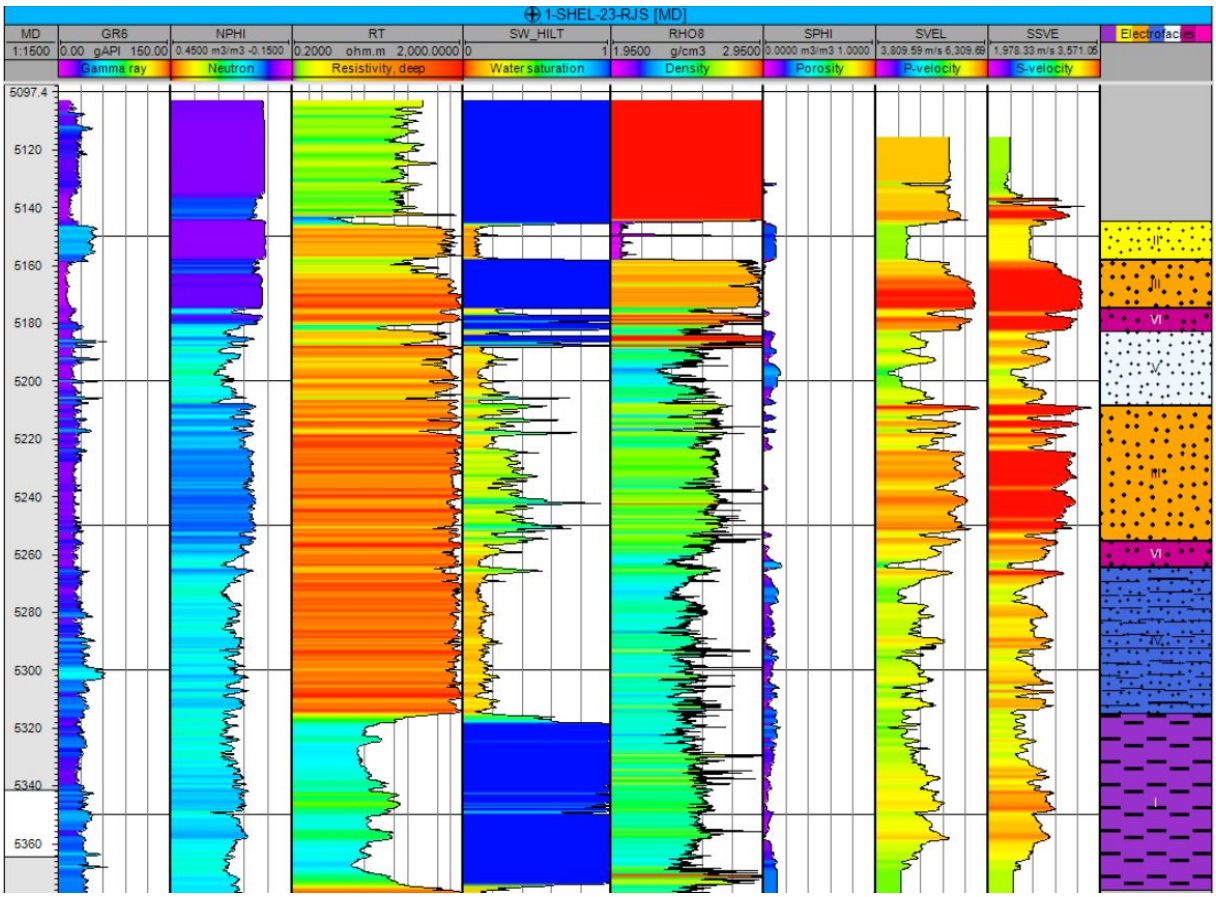



Figure 4.7: Manual electrofacies classication for the 1-SHEL-23-RJ well. Well log reference.

For unsupervised learning algorithms, due to the nature of algorithms, the flowchart is a little more simplified because algorithms does not need of a labeled and split data set. Thus, the research provided to the model the data set with the input features selected, depth matching, and statistical processing done. The flowchart for unsupervised learning can be seen in the figure (Figure 4.9). It is important to note that, despite of the unsupervised learning algorithms does not need of a data set labeled and split, the

final interpretation process it is more complicated because it is necessary to interpret the electrofacies created by the model. In supervised learning algorithms it is not necessary to interpret the electrofacies after the prediction, because the interpretation of the electrofacies is done previously in the manual classification stage.

The working flowchart applied in supervised learning algorithms can be seen in the figure 4.8, while the working flowchart for unsupervised learning algorithms can be visualized in the figure 4.9. In both working flowcharts, there are initial steps similar, they are: (i) data acquisition, (ii) features selection, (v) depth matching , and (vi) statistical processing. From this point the working flowchart of supervised and unsupervised learning algorithms are different because of the particularity of each algorithm. While in supervised learning algorithms the data set needs to be labeled and split into training and test data, in unsupervised learning algorithms these steps are not necessary.

### 4.4.1 Conventional workflow

The conventional process to classify electrofacies is made using a Geophysical well logs set able to represent the different rock properties in the subsurface. From the response of the Geophysical well logs, the geologist specialized begin the identification process of the electrofacies. To make the electrofacies classification, the geologist specialized verify the patterns present in the Geophysical well logs set and then individualize each of them. This process is repeated to all wells present in the petroleum field. If the petroleum field is huge and there are many wells in the field, then, the manual electrofacies classification can become a time-consuming process.

The manual classification shows to have unless two disadvantages, they are: (i) a time-consuming process, in the case that petroleum field has a lot of wells to be analyzed; (ii) the classification can change according to the specialized Geologist. In the second (ii) case the trouble is more serious because in this case, the electrofacies classification will not have a standard and consequently leads to one subjective interpretation of electrofacies.

### 4.4.2 Workflow in Supervised Learning algorithms

The working flowchart applied to supervised learning algorithms, Support Vector Machine and Neural Network can be visualized in the figure 4.8.

The steps that compound the working flowchart are: (i) data acquisition, (ii) features

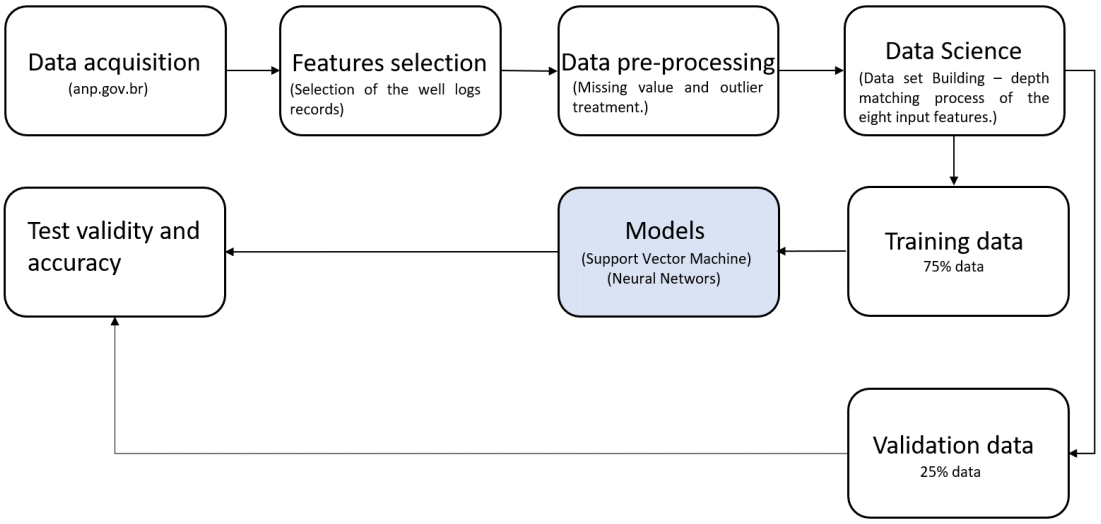



Figure 4.8: The flowchart of the proposed workflow applied in supervised learning algorithms.

selection, (iii) choice of the reference well, (iv) manual classification, (v) depth matching, (vi) statistical processing, (vii) data set splitting. After the data acquisition and features selection, the first step was to choose a reference well (1-SHEL-23-RJS) to work with the next steps. The choice of the reference well was made according to the number of records, number of electrofacies present in the well, and quality of the data present in the well. The well (1-SHEL-23-RJS) was chosen as the reference well because it had the best relationship among the number of records, electrofacies present and data quality. The fourth step (iv) adopted by the supervised learning flowchart (Figure 4.8) it is the same process that it is made for the conventional workflow described in the section 4.4.1. The manual electrofacies classification is necessary because it allows to labeled the data set of the reference well. Both algorithms utilized in the current work, Support Vector Machine and Neural Networks, require a labeled data set to train, learning, and making the predictions. The depth matching and statistical processing was called of data science stage by the research, these steps are the core of the research and were made and re-made with great care. Splitting the data set into training and test data it is necessary in supervised learning algorithms, because this step makes the statistical validation possible. From the labeled and processed data set, models built to predict electrofacies in new data sets (unknowns data sets) are able to learn. This happens because, when the reference well is labeled manually, the specialist Geologist it is teaching the algorithms

how the electrofacies classification must be done accurately to prevent big errors and lack of standards.

### 4.4.3 Workflow in Unsupervised Learning algorithms

The K-means workflow (Figure 4.9) is very similar to the workflow of the supervised learning algorithms (Figure 4.8), the main difference among them it is that the first uses an unlabeled data set, while the second needs a labeled data set. Furthermore, with unsupervised learning algorithms it is not possible the evaluate the error, once the data set is unlabeled, which makes it impossible to compare the results.

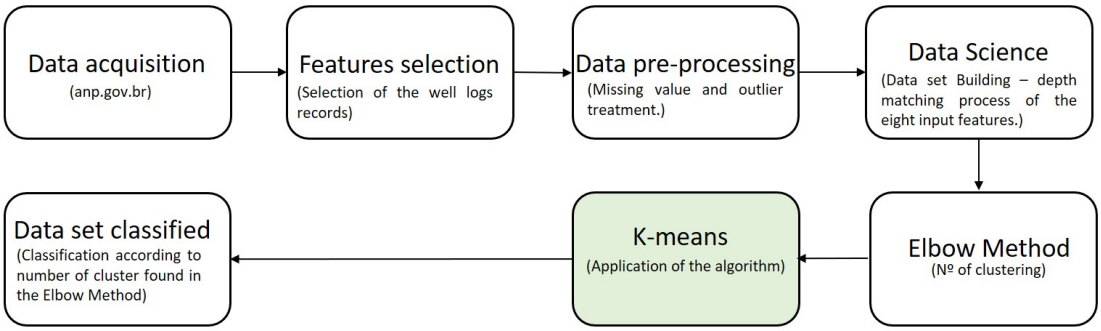



Figure 4.9: The flowchart of the proposed workflow applied in the unsupervised learning algorithms.

## 4.5 Statistical Validation

Estimating the accuracy of a classifier induced by supervised learning algorithms is important not only to predict its future prediction accuracy, but also for choosing the best classifier from a given set of models [Koh95] [Yan07]. In the work was used the confusion matrix [HTF01], [JWHT13], [JS11] and k-Fold Cross-Validating [BG04], [AGG+12] to evaluate the model, [AAV+21]. The confusion matrix is constructed by plotting the predicted and real data in the axis x and y of the table, in a multi-class confusion matrix. The element $N_{i,j}$ present in the confusion matrix is called *true positive*, situation when the predicted data is equal to the real data, this situation is verified when the row i is equal to the column j ($i = j$). There are four elements in the confusion matrix responsible for statistical validation, they are accuracy, precision, recall, and F1 score. The accuracy

(Eq.4.5.1) can be defined mathematically as:

$$\textbf{Accuracy} = \frac{\sum_{i=1}^{C}(TP_i + TN_i)}{\sum_{i=1}^{C} N_i} \tag{4.5.1}$$

where $TP_i$ are the true positives, the positive data that were correctly predicted in the class $C_i$, C is the number of classes, $TN_i$ are the true negatives, and $N_i$ is the number of samples in class $C_i$.

The precision (Eq.4.5.2) evaluates among all samples predicted as positive by the model how many are true positive to the class predicted. The precision evaluation can be understood as the number of true positive that were correctly predicted as positive in the set of all data predicted as positive. The precision evaluation can be described as:

$$\textbf{Precision}(C_i) = \frac{TP_i}{TP_i + FP_i} \tag{4.5.2}$$

where $C_i$ indicates the class that measure was taken. $FP_i$ represents the false positive, the samples that were predicted to be positive, but in fact, are negative.

The recall (Eq.4.5.3) evaluates among all samples that are positive how many were predicted as positive. The recall evaluation can be understood as the number of true positive that were correctly predicted as positive in the set of positive data. The recall is defined as:

$$\textbf{Recall}(C_i) = \frac{TP_i}{TP_i + FN_i} \tag{4.5.3}$$

where $FN_i$ are the false negative, samples that were predicted to be false, but in fact are positive.

The F1-score (Eq.4.5.4) is a combination between precision and recall. Mathematically the F1-score can be understood as a harmonic mean of recall and precision and, in some cases, can be considered a measure more representative than accuracy. The F1-Score is defined as:

$$\textbf{F1-Score}\ (C_i) = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4.5.4}$$

The other technique used to evaluate the model performance was the k-Fold Cross-Validating, the main use of this technique is to verify the capacity of generalization of the model and to estimate the prediction accuracy [BG04], [Fus11]. This method consists of splitting the dataset into k subsets of equal or nearly equal sizes, where each subset is known as a fold and is stratified: that is, each fold attempts to retain the same class distribution. A model is then trained with k - 1 folds and tested on the remaining fold. This process is repeated k times until all sets have been used for testing once. Thus, the experiment returns k estimates of the model classification error [AAV+21].

## 4.6 Vertical Heterogeneity

From the products obtained by models I and II was created the index of heterogeneity (Eq.4.6.1) of reservoirs zones. This parameter was created to measure the heterogeneity level of the reservoir zone predicted by models I and II,

$$\mathbf{h} = \frac{TTN}{TTR + TTN} \tag{4.6.1}$$

where, $TTN$ is the total thickness of non reservoir zone, $TTR$ is the total thickness of reservoir zone, and $h$ is the level of heterogeneity.

# Chapter 5

# Results

The chapter is built showing all results derived from the current research on articles and preprints, thus in this chapter, the reader will be directed to the respective articles and preprints in the appendices section. The section 5.1 presents the results obtained with the model I - Support Vector Machine. In the section 5.2 are demonstrates the results about the model II - Neural Networks. From section 5.3 it is possible to visualized the results about the vertical heterogeneity in both models. The section 5.4 illustrates the first product generate by the research using unsupervised learning algorithms. From section 5.5 the reader will know the first work developed with supervised learning algorithms, and the section 5.6 shows the latest work done in the current research. In addition, the chapter also discusses the main results achieved with the research.

## 5.1 Model I - Support Vector Machine

From the techniques illustrated in the statistical validation section 4.5, results about the performance of the models applied in the research were obtained. In the model I (SVM), the confusion matrix and classification report are shown in tables 5.1 and 5.2:

From the table 5.1, it is possible to analysed the performance of the model I in the test data set. According to table 5.1, there are 406 records in the test data set, considering that the total number of records in the data set is 1617, then the test data set corresponds to 25.1% of the total data set. When analyzing the rows and columns of the confusion matrix, it is possible to notice that the classification error is greater in rows 5 and 3 than in the others. The consequence is that, rows 5 and 3 that corresponds to electrofacies 2

Table 5.1: SVM - Confusion Matrix

| | | **Predict** | | | |
|---|---|---|---|---|---|
| **Real** | **123** | 0 | 0 | 2 | 0 |
| | 0 | **25** | 0 | 0 | 0 |
| | 0 | 1 | **113** | 7 | 4 |
| | 0 | 0 | 2 | **85** | 0 |
| | 1 | 0 | 8 | 3 | **32** |

and 4 in the table 5.2 presents the first and second lowest values of F1-score, respectively. The classification report of the model I, table 5.2, presents the metrics of each electrofacies predicted by the model I. The best electrofacies performance it is found in eletrofacies 0 followed by electrofacies 1. The worst electrofacies performance can be visualized in eletrofacies 4 followed by electrofacies 2. The classification report allows to visualized that just electrofacies 4 has F1-score bellow to 90%. The model I has presented a hit rate in the training data set of 93.7%, and a value of 92.9% in the test data set. The mean final accuracy found by the Confusion matrix method for model I was about 93%.

Table 5.2: SVM - Classification Report

| **EF** | **Precision** | **Recall** | **F1-score** | **Spp** |
|---|---|---|---|---|
| 0 | 0.99 | 0.98 | 0.99 | 125 |
| 1 | 0.96 | 1.00 | 0.98 | 25 |
| 2 | 0.92 | 0.90 | 0.91 | 125 |
| 3 | 0.88 | 0.98 | 0.92 | 87 |
| 4 | 0.89 | 0.73 | 0.80 | 44 |
| **Acurracy = 0.93** | | | | |

where, $EF = electrofacies$ and $Spp = support$, the number of samples that were used in the evaluation.

It is important to notice the improvement of the model I during research development. Initially, the model I was development according to a manual electrofacies classification based on 6 electrofacies (Figure 4.7), and trained with 8 input features, these results can be visualized in section 5.5. During the research development was noted that the model

I was very confused on distinguish the electrofacies II (orange) and VI (pink), figure 4.7. Thus, the electrofacies II (orange) and VI (pink) were merged due to their similar values of input features. Beyond that, according to analysis of correlation matrix (Figures 4.6 and 4.5) the P-velocity input feature was removed in order to optimize the model. These changes were responsible to improve the final accuracy of the model I in 7%. The results illustrated in this section and in section 5.6 are derived from these changes. The well log section with the electrofacies predicted by the model I can be visualized in the figure 5.1.

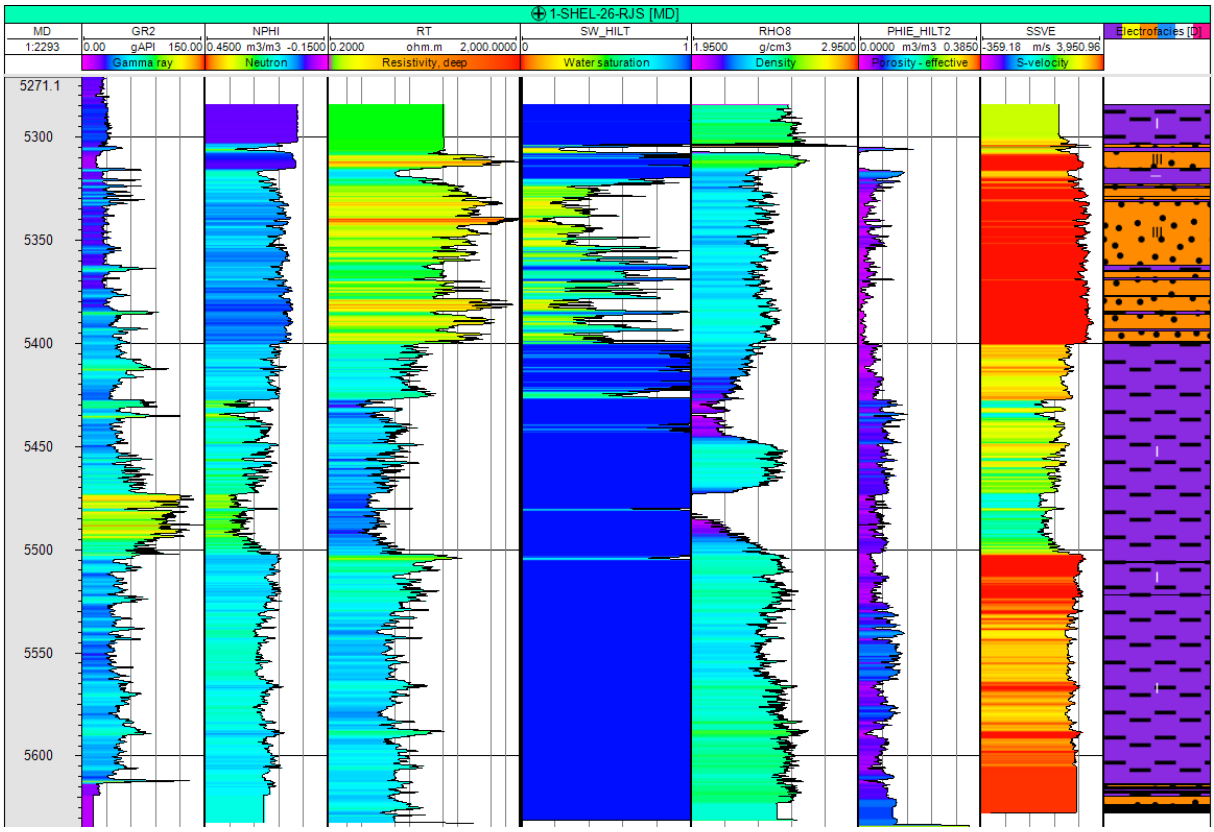



Figure 5.1: Electrofacies predicted by the model I (Support vector machine) for the well 1-SHEL-26-RJ.

In relation to the K-fold Cross-Validation method (Table 5.5), it is important to notice that the model I presents a lower final accuracy than confusion matrix method. In addition, the model I has a standard deviation of approximately 10%. This high standard deviation value indicates that, according to the slice of data set, the performance of the model can imply a large variation in the final accuracy.

## 5.2 Model II - Neural Networks

The model II was development based on Neural Networks algorithms and demonstrates better results than model I. From the beginning, the model II was development according to a manual electrofacies classification based on 5 electrofacies, and trained with 7 input features, these results can be visualized in section 5.6. From the moment that the model I was well calibrated, then the model II was created. The main goal in creation of the model II was to verify the convergence of the models in predicted electrofacies. Secondarily, the performance comparison between models I and II was made. In relation with the model I, the model II obtained 96.6% of final accuracy evaluation by the confusion matrix method (Tables 5.3 and 5.4), and 96.9% of final accuracy evaluation by the K-fold Cross-Validating (Table 5.5). These results of the model II when compared with the same results of the model I were responsible to represent 3.7% of increase in the final accuracy evaluation by the confusion matrix method, and 10% of increase in final accuracy evaluation by the K-fold Cross-Validating method. The execution time in model II was 41.45s, while in the model I was 0.096s. The results are illustrated in this section and in section 5.6. The confusion matrix and classification report generate by the Neural Network model is shown in Tables 5.3 and 5.4.

Table 5.3: Neural Network - Confusion Matrix

| | | **Predict** | | | | |
|---|---|---|---|---|---|---|
| **Real** | | **143** | 0 | 0 | 1 | 0 |
| | | 0 | **24** | 0 | 0 | 0 |
| | | 0 | 0 | **119** | 1 | 0 |
| | | 1 | 0 | 5 | **68** | 0 |
| | | 2 | 0 | 3 | 0 | **39** |

In the K-fold Cross-Validation method, the model II has demonstrated a consitency in electrofacies prediction almost 10x higher than model I. In the model II the standard deviation it is approximately 1% and has a final accuracy evaluation in the K-fold Cross-Validation method (96.9%) very similar to the final accuracy evaluation found in the Confusion matrix method (96.6%). Opposite to the model I, the low standard deviation in model II indicates a high consistency in the electrofacies predicted by the model. The

Table 5.4: Neural Network - Classification Report

| EF | Precision | Recall | F1-score | Spp |
|---|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.99 | 144 |
| 1 | 1.00 | 1.00 | 1.00 | 24 |
| 2 | 0.94 | 0.97 | 0.96 | 120 |
| 3 | 0.96 | 0.93 | 0.95 | 74 |
| 4 | 0.97 | 0.89 | 0.93 | 44 |
| **Acurracy = 0.97** | | | | |

electrofacies predicted by the model II can be visualized in the figure 5.2.

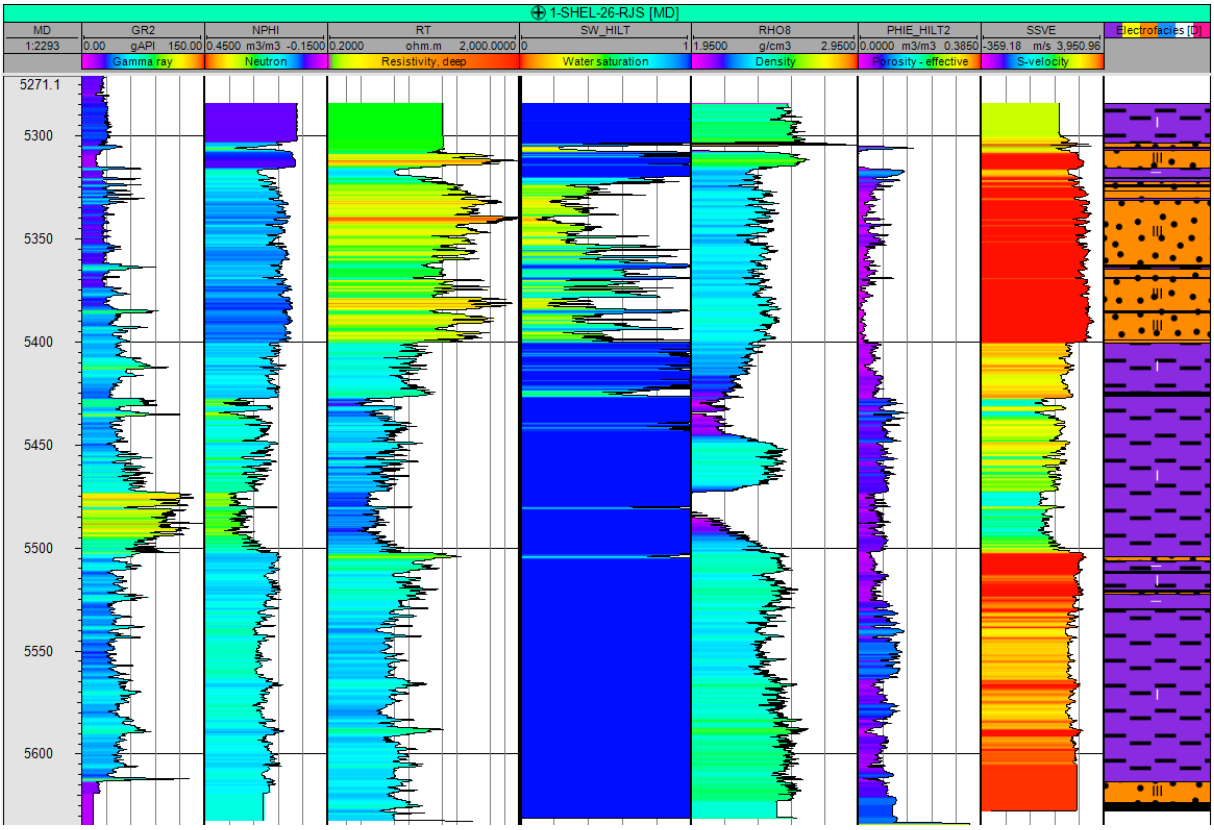



Figure 5.2: Electrofacies predicted by the model II (Neural Networks) for the well 1-SHEL-26-RJ.

It is possible to observe in tables 5.2 and 5.4 an increase in the final value of all statistical parameters in Neural Network model when compared to the SVM model. Final accuracy, F1 score, precision, and recall are higher for all electrofacies present in the work. These results revealed a greater robustness of Neural Networks algorithms and a better

Table 5.5: Supervised learning Models

| **Accuracy evaluation k-Fold Cross-Validating** | |
|---|---|
| Model I (SVM) | Model II (Neural Network) |
| 0.8307 | 0.9599 |
| 0.8492 | 0.9692 |
| 0.9661 | 0.9692 |
| 0.9507 | 0.9846 |
| 0.7314 | 0.9629 |
| **Mean ± Std** | **Mean ± Std** |
| 0.8656 ± 0.0959 | 0.9692 ± 0.0095 |

matching with the data set of the research.

When AI algorithms are working on a data set, it is important to check that the algorithms are not in an overfitting or underfitting case. The overfitting case is verified when the algorithm presents a great performance in the training data set, but when submitted to the test data set performs poorly.

The model II has presented hit rate in training and test data set of 96.9% and 96.6%.

In both models, the underfitting and overfitting of models were verified and, in both cases, the hit rate found in the training data set was very similar with the hit rate in the test data set. This verification process indicates that the algorithms of the models are not in an overfitting case. The case of underfitting occurs when the algorithm performs poorly on the training and test data set. As has been shown with the statistical results, the algorithms applied in SVM model and Neural Network model are not in the case of underfitting.

## 5.3 Vertical Heterogeneity - Models I and II

An extra and valuable information that can be obtained from the electrofacies prediction made by models I and II it is the evaluation of the vertical heterogeneity/homogeneity of the reservoir zones. The research used the equation 4.6.1 to evaluate the vertical het-

erogeneity.

The Model I presents 13 sections of non reservoir zone inside the reservoir zone. The thirteen sections thickness of model I can be visualized in the Table 5.6. The minimum and maximum thickness of non reservoir zone inside of reservoir zone are 0.15 m, and 7.01 m respectively. The mean thickness of the non reservoir zone inside the reservoir zone to the Model I it is 1.20 meters. The Model II presents 14 sections of non reservoir zone inside the reservoir zone. The fourteen sections thickness of the model II can be seen in the Table 5.7, it is possible to note the minimum and maximum thickness 0.15 m, and 4.42 m respectively. From the Table 5.7 it is possible to calculate the mean thickness of the non reservoir zone inside the reservoir zone to the Model II, 1.04 meters.

Table 5.6: Sections of non reservoir zones - Model I

| Depth Interval (m) | Section thickness (m) |
|---|---|
| 5304.74 - 5304.89 | 0.15 |
| 5305.20 - 5307.18 | 1.98 |
| 5315.41 - 5322.42 | 7.01 |
| 5323.64 - 5324.55 | 0.91 |
| 5328.51 - 5328.67 | 0.16 |
| 5330.34 - 5330.95 | 0.61 |
| 5362.35 - 5364.48 | 2.13 |
| 5364.79 - 5365.09 | 0.30 |
| 5368.44 - 5369.36 | 0.92 |
| 5376.98 - 5377.28 | 0.30 |
| 5384.60 - 5385.97 | 1.37 |
| 5392.98 - 5393.74 | 0.76 |
| 5398.92 - 5399.23 | 0.31 |
| Total | 16.90 |

The table 5.8 shows the level of vertical heterogeneity in the well log section predicted

Table 5.7: Sections of non reservoir zones - Model II

| Depth Interval (m) | Section thickness (m) |
|---|---|
| 5305.35 - 5306.87 | 1.52 |
| 5315.56 - 5319.98 | 4.42 |
| 5320.44 - 5322.27 | 1.83 |
| 5323.48 - 5324.55 | 1.07 |
| 5326.53 - 5326.84 | 0.31 |
| 5328.51 - 5328.67 | 0.16 |
| 5330.19 - 5330.95 | 0.76 |
| 5353.05 - 5353.32 | 0.27 |
| 5362.65 - 5364.48 | 1.83 |
| 5364.94 - 5365.09 | 0.15 |
| 5368.60 - 5369.20 | 0.60 |
| 5376.98 - 5377.13 | 0.15 |
| 5384.75 - 5385.97 | 1.22 |
| 5392.98 - 5393.28 | 0.30 |
| Total | 14.59 |

by the Model I. And the table 5.9 shows the level of vertical heterogeneity in the well log section predicted by the Model II.

Models I and II has similar level of vertical heterogeneity. The difference of 2% between the vertical heterogeneity in models reinforce the convergence of the predicted electrofacies by models I and II.

Table 5.8: Heterogeneity level of reservoir zone in the Model I

| | |
|---|---|
| Total Thickness of reservoir zone | 96.62 m |
| Total Thickness of non-reservoir zone | 16.90 m |
| Nº of sections non reservoir zone | 13 |
| Vertical heterogeneity of the reservoir zone | 17.49% |

Table 5.9: Heterogeneity level of reservoir zone in the Model II

| | |
|---|---|
| Total Thickness of reservoir zone | 96.91 m |
| Total Thickness of non-reservoir zone | 14.59 m |
| Nº of sections non reservoir zone | 14 |
| Vertical heterogeneity of the reservoir zone | 15.06% |

## 5.4 K-means algorithm approach for automate the electrofacies classification: An exploratory study ap-plied in Brazilian Pre-Salt, Santos Basin A

## 5.5 Applying supervised machine learning model to classify electrofacies in a Brazilian Pre-salt well-bore B

## 5.6 An approach by Neural Networks and Support Vector Machine in classification and prediction of carbonates electrofacies C

# Chapter 6

# Final Considerations

The last chapter will go to talk about the final considerations about the research, after 2 years of work, this chapter has the main goal to finish the master's degree chapter and opening the P.hD. chapter. At the beginning of this chapter, the author's feelings are of relief and satisfaction with the work developed. Section 6.1 shows the discussion about the results obtained in the research. While the section 6.2 the author makes the conclusions about the research.

## 6.1 Discussion

The research brought to the light of the comprehension three different models to classify and predict electrofacies from well logs. The first model presented by the research was the k-means, an unsupervised learning algorithm. The k-means showed a high potential to cluster data set from well log. An advantage of k-means over other algorithms applied in the research is that k-means does not need to be taught because the algorithm clustering the data set from the data set distribution. On the other hand, the k-means also has disadvantages, the main is due to in many cases the pattern found does not have meaning in the real world. The SVM and Neural Networks are supervised learning algorithms and with the help of the human knowledge has been shown a good alternative to automate the classification process of electrofacies. The results of the research has demonstrated the better performance of Neural Networks in relation the SVM.

According to the confusion matrix (Table 5.1) and classification report (Table 5.2) tables, it is possible to notice that the SVM algorithm has the biggest statistical errors

associated with the prediction of the electrofacies 04 (Table 5.2), row 5 (Table 5.1). The confusion matrix (Table 5.1) shows that out of 44 samples, the SVM algorithm failed in 12 predictions. Among the 12 failed predictions, more than half(8) were classified as electrofacies 02. These results have a direct impact on the classification report, causing the statistical scores of precision, recall, and f1-score to drop significantly in relation to the statistical scores of other electrofacies. One way to try to improve the statistics of the algorithm is to begin with the range of values of the parameters that make up the electrofacies. An alternative, in the case where the detail level can be decreased, is to merge the electrofacies that the algorithm is confusing in the prediction process, electrofacies 02 and 04. This process can increase the statistical scores, but on the other hand, the detail level of the electrofacies description decrease. In the current research, final values of hit rate above 90% were considered acceptable, thus was not necessary to merge the electrofacies and consequently to lose the current level of detail.

In Neural Networks model, the confusion matrix (Table 5.3) presents a matching between the predicted and real electrofacies considerably greater than the confusion matrix of the SVM model (Table 5.1). It is possible to verify by means of the classification report (Table 5.4) that all statistical scores have increased in Neural Network model. This occurs due to the better matching between the algorithm and the research data set, and also because the Neural Network is a more robust algorithm than SVM. In the Neural Networks model (Table 5.3), the problem of confusion between the prediction of electrofacies 02 and 04 can be ignored, because of 44 samples labeled as electrofacies 04, only three were predicted as electrofacies 02.

The well logs sections of SVM model and Neural Networks model (Appendix C) are very similar, among the five electrofacies available to fill the section, the models predicted the presence of only two, electrofacies 0 (purple) and 02 (orange). In the context of Gato do Mato oil field, these electrofacies presents two distinct zones, reservoir zone and non-reservoir zone. The electrofacies 0, due to the greater content of clay, verified by the gamma rays records and high values of density and water saturation represents a section with properties of non-reservoir zone. On the other hand, the electrofacies 02 have low values of gamma rays and water saturation, high values of resistivity, and moderate values of density, indicating properties of a reservoir zone.

In the well log 1-SHEL-26-RJS, the most part is composed of non-reservoir zones

(electrofacies 0), and a small portion is filled with reservoir zone(electrofacies 02). This behavior in the classification, illustrated by models SVM and Neural Networks (Appendix C), can be explained by the Gamma Rays. It is possible to observe that the values of Gamma Rays are high in most part of the well, and the input parameter, Gamma Rays, is one of the most essential parameters used to distinguish between argilous zones and non-argilous zones and this zones that will auxiliary in the classification of reservoir and non-reservoir zones. Thus, the main reason why the 1-SHEL-26-RJS well is mostly classified as a non-reservoir zone is because of the values of Gamma rays that lead the algorithms of SVM and Neural Network to classify these portions as non-reservoir zones(electrofacies 0).

The main differences between the well log sections predicted by the SVM model and Neural Network model are that, the well log section predicted by the SVM model is more heterogeneous than well log section predicted by the Neural Network model. In other words, the SVM model has an electrofacies variation higher than Neural Network model, and in the Gato do Mato oil field context, the Neural Network model represents a behavior closer to reality than SVM model.

There are differences between the model applied in the Rio oil & gas using the SVM algorithm and the model applied in appendix C. The models have generated different hit-rate and results. In the article submitted to the Rio oil & gas, the electrofacies pink and orange are not merged, opposite to model applied in appendix C where the electrofacies pink and orange are merged. This was made because from the confusion matrix was possible to note that the algorithm had many difficulties in differentiating the electrofacies pink and orange, and consequently the precision was negatively affected decreasing the final value. This happened because the geophysical well logs of the electrofacies orange and pink were very close.

According to results presented in models SVM and Neural Network, it is clear that both models reaching a high-performance level. Both algorithms are capable to obtain a hit rate greater than 90%, therefore both models can be used in electrofacies predictions. When a performance comparison is made between the models, the Neural Network model shows to be 3.7% better than the SVM model. In the well log section it is possible to observe that the classification made by the models are very similar, but in the case of the well log section produced by the Neural Network model, the electrofacies predicted

shows to be smoother and has smaller electrofacies variation than the SVM model. This behavior observed in the Neural Network model is closer to the natural behavior verified in the oil field Gato do Mato, located in the Brazilian pre-salt.

## 6.2 Conclusions

The research has been shown progress to understand and improve the models applied in research, until the moment, gains in final accuracy, precision, recall and F1-score has been made in the supervised learning models, and has shown the right way to follow. The results obtained with the supervised learning models can be considered promising once that the algorithms applied in models showed to be capable to predict with high precision the data set. In other words, the supervised learning models shows a high potential to automate the process of electrofacies classification. This automation, if well calibrated as can be seen in model of Neural Network applied in the research can save time and resources for petroleum companies in large oil fields with a lot of well logs information.

The k-means algorithm also has shown a high potential once that the model does not need a human to teach the data patterns. But after the classification process, human knowledge it is necessary to interpret the classification made by the model. In supervised learning is different, because initially, we need to teach the model, but after the teaching process, the data set classified does not need to be interpreted.

An advantage of models applied in the research to predict and classify the electrofacies is that the algorithms can quickly analyze several features, and thus, to predict patterns in regions difficult to be find by human eyes. Another advantage is that, in supervised learning models, the input features, once associated with the corresponding electrofacies, can be used to generalize the overall reservoir. Therefore, it becomes possible the data set that represents a few inches of the surrounding well can be interpolated by the wells in the oil field and, consequently, to generate a model of the entire reservoir.

The current research approached themes of high importance in the modern scenario, the union between well logs data set and AI algorithms have showed a success combination. Therefore, the work contribute for the Geoscience and AI literatures as a practical case of AI algorithms automating an important process in the petroleum geology.

The models presented in the research open doors to other applications in the O&G

sector. The own models applied in the current research can be recalibrated for predictions of others sectors. Once is known the point where we want to arrive, the models presented in the research can be recalibrated to reach similar or better performance. There are researches being developed by the research group with the main goal to explore other possibilities to apply the models developed in the current research in other areas of O&G. The future of O&G and other technological sectors has shown that the combination of AI algorithms and human knowledge is a long-term marriage, this seems clear at the present moment in human development because a few years earlier, the processing of the human brain, was in many cases, capable of working with the amount of data for processing, learning, and teaching, but at the present moment, the amount of data has become so huge that is impossible for the human brain to process all relationship and implications of the collected data. Therefore, it is possible to affirm that the collaboration between AI algorithms and humans will remain strong for long years.

# Bibliography

[AAV+21] Carlos EM dos Anjos, Manuel RV Avila, Adna GP Vasconcelos, Aurea M Pereira Neta, Lizianne C Medeiros, Alexandre G Evsukoff, Rodrigo Surmas, and Luiz Landau, *Deep learning for lithological classification of carbonate rock micro-ct images*, Computational Geosciences (2021), 1–13. ↑64, 66

[AB09] Martin Anthony and Peter L Bartlett, *Neural network learning: Theoretical foundations*, cambridge university press, 2009. ↑18

[AGG+12] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella, *The'k'in k-fold cross validation.*, Esann, 2012, pp. 441–446. ↑64

[AKG04] George B Asquith, Daniel Krygowski, and Charles R Gibson, *Basic well log analysis*, Vol. 16, American Association of Petroleum Geologists Tulsa, 2004. ↑23, 44

[Ara14] Mitsuru Arai, *Aptian/albian (early cretaceous) paleogeography of the south atlantic: a paleontological perspective*, Brazilian Journal of Geology **44** (2014), no. 2, 339–350. ↑20, 30

[AW99] Shun-ichi Amari and Si Wu, *Improving support vector machine classifiers by modifying kernel functions*, Neural Networks **12** (1999), no. 6, 783–789. ↑38

[Ba13] Lei Ba, *Adaptive dropout for training deep neural networks*, Ph.D. Thesis, 2013. ↑42

[BA14] Purnima Bholowalia and Kumar Arvind, *Ebk-means: A clustering technique based on elbow method and k-means in wsn*, International Journal of Computer Applications 105(9), 2014. ↑36

[BCHC09] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, *Pearson correlation coefficient*, Noise reduction in speech processing, 2009, pp. 1–4. ↑57

[BDL20] Digvijay Boob, Santanu S Dey, and Guanghui Lan, *Complexity of training relu neural network*, Discrete Optimization (2020), 100620. ↑42

[Ben09] Yoshua Bengio, *Learning deep architectures for ai*, Now Publishers Inc, 2009. ↑19

[BG04] Yoshua Bengio and Yves Grandvalet, *No unbiased estimator of the variance of k-fold cross-validation*, Journal of machine learning research **5** (2004), no. Sep, 1089–1105. ↑64, 66

[BG05] Irad Ben-Gal, *Outlier detection*, Data mining and knowledge discovery handbook, 2005, pp. 131–146. ↑58

[BGV] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik, *A training algorithm for optimal margin classifiers*, Proceedings of the 5th annual acm workshop on computational learning theory, pp. 144–152. ↑38

[Big92] Ed L Bigelow, *Introduction to wireline log analysis*, Western Atlas International, 1992. ↑12, 45

[Bis06] Christopher M Bishop, *Pattern recognition*, Machine learning **128** (2006), no. 9. ↑38

[CA16] M Emre Celebi and Kemal Aydin, *Unsupervised learning algorithms*, Springer, 2016. ↑34

[CAC+08] Hung Kiang Chang, Mario Luis Assine, Fernando Santos Corrêa, Julio Setsuo Tinen, Alexandre Campane Vidal, and Luzia Koike, *Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na bacia de santos*, Revista Brasileira de Geociências **38** (2008), no. 2 suppl, 29–46. ↑19, 27, 30

[CC10] Denis Cousineau and Sylvain Chartier, *Outliers detection and treatment: a review.*, International Journal of Psychological Research **3** (2010), no. 1, 58–67. ↑55, 58

[CL11] Chih-Chung Chang and Chih-Jen Lin, *Libsvm: a library for support vector machines*, ACM transactions on intelligent systems and technology (TIST) **2** (2011), no. 3, 1–27. ↑38

[Cop04] Ben Coppin, *Artificial intelligence illuminated*, Jones & Bartlett Learning, 2004. ↑33

[CS01] Koby Crammer and Yoram Singer, *On the algorithmic implementation of multiclass kernel-based vector machines*, Journal of machine learning research **2** (2001), no. Dec, 265–292. ↑38

[CV95] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine learning **20** (1995), no. 3, 273–297. ↑37, 38

[CWG08] Mario Carminatti, Breno Wolff, and Luiz Gamboa, *New exploratory frontiers in brazil*, 19th world petroleum congress, 2008. ↑19

[Dav18] John C Davis, *Electrofacies in reservoir characterization*, Handbook of mathematical geosciences, 2018, pp. 211–223. ↑19

[Daw11] Robert Dawson, *How significant is a boxplot outlier?*, Journal of Statistics Education **19** (2011), no. 2. ↑55, 56

[DFK+99] Petros Drineas, Alan M Frieze, Ravi Kannan, Santosh Vempala, and V Vinay, *Clustering in large graphs and matrices.*, Soda, 1999, pp. 99, 291–299. ↑35

[EDFK10] Tristan Euzen, Eric Delamaide, Tom Feuchtwanger, and Kim D Kingsmith, *Well log cluster analysis: an innovative tool for unconventional exploration*, Canadian unconventional resources and international petroleum conference, 2010. ↑18

[ES07] Darwin V Ellis and Julian M Singer, *Well logging for earth scientists*, Vol. 692, Springer, 2007. ↑23, 44, 48

[FA19] Hatem A Fayed and Amir F Atiya, *Speed up grid-search for parameter selection of support vector machines*, Applied Soft Computing **80** (2019), 202–210. ↑37, 40

[Fan10] John Fanchi, *Integrated reservoir asset management: principles and best practices*, Gulf Professional Publishing, 2010. ↑46

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, *Liblinear: A library for large linear classification*, the Journal of machine Learning research **9** (2008), 1871–1874. ↑38

[FI05] Frauke Friedrichs and Christian Igel, *Evolutionary tuning of multiple svm parameters*, Neurocomputing **64** (2005), 107–117. ↑40

[Flo13] Romeo M Flores, *Coal and coalbed gas: fueling the future*, Newnes, 2013. ↑47

[FS16] John R Fanchi and John P Seidle, *Multiphase flow in porous media*, Multiphase flow handbook, 2016, pp. 685–728. ↑46, 47, 48, 50

[Fus11] Tadayoshi Fushiki, *Estimation of prediction error by using k-fold cross-validation*, Statistics and Computing **21** (2011), no. 2, 137–146. ↑66

[Gar12] Savio Francis De Melo Garcia, *Restauração estrutural da halotectônica na porção central da bacia de santos e implicações para os sistemas petrolíferos.* (2012). ↑27

[GG93] Stephen I Gallant and Stephen I Gallant, *Neural network learning and expert systems*, MIT press, 1993. ↑18

[Gha03] Zoubin Ghahramani, *Unsupervised learning*, Summer school on machine learning, 2003, pp. 72–112. ↑34, 36

[Glo00] Paul WJ Glover, *Petrophysics*, University of Aberdeen, UK (2000). ↑50

[GMDS+08] LAP Gamboa, MA Pinheiro Machado, DP Da Silveira, JTR De Freitas, SR Pereisa Da Silva, W Mohriak, P Szatmari, and S Anjos, *Evaporitos estratificados no atlântico sul: interpretação sísmica e controle tectono-estratigráfico na bacia de santos*, Sal: Geologia e Tectônica, Exemplos nas Basicas Brasileiras (2008), 340–359. ↑31

[HPS05] Sariel Har-Peled and Bardia Sadri, *How fast is the k-means method?*, Algorithmica **41** (2005), no. 3, 185–202. ↑18

[HRYB98] David C Howell, M Rogier, V Yzerbyt, and Y Bestgen, *Statistical methods in human sciences*, New York: Wadsworth **721** (1998). ↑54

[HSK+12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, arXiv preprint arXiv:1207.0580 (2012). ↑43

[HSM+00] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung, *Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit*, Nature **405** (2000), no. 6789, 947–951. ↑42

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola, *Kernel methods in machine learning*, The annals of statistics (2008), 1171–1220. ↑38

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning. springer series in statistics*, :, 2001. ↑64

[HTF09] ———, *Unsupervised learning*, The elements of statistical learning, 2009, pp. 485–585. ↑33, 34, 36

[Jai10] Anil K Jain, *Data clustering: 50 years beyond k-means*, Pattern recognition letters **31** (2010), no. 8, 651–666. ↑18, 34

[JCGJ17] Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou, *Efficient softmax approximation for gpus*, International conference on machine learning, 2017, pp. 1302–1310. ↑42

[JF01] AH Joarder and M Firozzaman, *Quartiles for discrete data*, Teaching Statistics **23** (2001), no. 3, 86–89. ↑55

[JM16] Paulo Roberto Schroeder Johann and Rubens Caldeira Monteiro, *Geophysical reservoir characterization and monitoring at brazilian pre-salt oil fields*, Offshore technology conference, 2016. ↑19

[JS11] Nathalie Japkowicz and Mohak Shah, *Evaluating learning algorithms: a classification perspective*, Cambridge University Press, 2011. ↑64

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *Statistical learning*, An introduction to statistical learning, 2013, pp. 15–57. ↑64

[KB14] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014). ↑43

[Kit77] Bill A Kithas, *Lithology, gas detection, and rock properties from acoustic logging systems*, Spe oklahoma city regional meeting, 1977. ↑49

[KK06a] Bishnu Kumar and Mahendra Kishore, *Electrofacies classification–a critical approach*, 6th international conference & exposition on petroleum geophysics, new delhi, india, 2006, pp. 822–825. ↑19, 23

[KK06b] ———, *Electrofacies classification—a critical approach*, 6th international conference and exposition on petroleum geophysics, kolkata, india, 2006, pp. 822–825. ↑19

[KM13] Trupti M Kodinariya and Prashant R Makwana, *Review on determining number of cluster in k-means clustering*, International Journal, 1(6), 90-95, 2013. ↑36

[KMN+02] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu, *An efficient k-means clustering algorithm: Analysis and implementation*, IEEE transactions on pattern analysis and machine intelligence **24** (2002), no. 7, 881–892. ↑34

[KMP05] Miroslav Kobr, Stanislav Mareš, and Frederick Paillet, *Geophysical well logging*, Hydrogeophysics, 2005, pp. 291–331. ↑23

[Koh95] Ron Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, 1995, pp. 1137–1145. ↑64

[KSIN15] Hamed Kiaei, Yousef Sharghi, Ali Kadkhodaie Ilkhchi, and Mehrangiz Naderi, *3d modeling of reservoir electrofacies using integration clustering and geostatistic method in central field of persian gulf*, Journal of Petroleum Science and Engineering **135** (2015), 152–160. ↑34

[KVLD12] Michelle Chaves Kuroda, Alexandre Campane Vidal, Emilson Pereira Leite, and Rodrigo Duarte Drummond, *Electrofacies characterization using self-organizing maps*, Brazilian Journal of Geophysics **30** (2012), no. 3. ↑18, 23

[KZ00] Stephen Kokoska and Daniel Zwillinger, *Crc standard probability and statistics tables and formulae*, Crc Press, 2000. ↑54

[KZP07] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas, *Supervised machine learning: A review of classification techniques*, Emerging artificial intelligence applications in computer engineering **160** (2007), no. 1, 3–24. ↑18

[LJK+00] Jorma Laurikkala, Martti Juhola, Erna Kentala, N Lavrac, S Miksch, and B Kavsek, *Informal identification of outliers in medical data*, Fifth international workshop on intelligent data analysis in medicine and pharmacology, 2000, pp. 20–24. ↑55

[LS18] Guifang Lin and Wei Shen, *Research on convolutional neural network based on improved relu piecewise activation function*, Procedia computer science **131** (2018), 977–984. ↑42

[LSJ04] Hancong Liu, Sirish Shah, and Wei Jiang, *On-line outlier detection and data cleaning*, Computers & chemical engineering **28** (2004), no. 9, 1635–1647. ↑58

[Mac67] James MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the fifth berkeley symposium on mathematical statistics and probability, 1967, pp. 281–297. ↑18

[MHW18] D[hendra] Marutho, S[unarna] H[endra] Handaka, and E[kaprana] Wijaya, *The determination of cluster number at k-mean using elbow method and purity evaluation on headline news*, 2018 International Seminar on Application for Technology of Information and Communication, IEEE. 533–538., 2018. ↑36

[Mia10] Andrew D Miall, *The geology of stratigraphic sequences*, Springer Science & Business Media, 2010. ↑19

[Mil91] Jeff Miller, *Reaction time analysis with outlier exclusion: Bias varies with sample size*, The quarterly journal of experimental psychology **43** (1991), no. 4, 907–912. ↑53

[MMGM07] Jobel Lourenço Pinheiro Moreira, Cláudio Valdetaro Madeira, João Alexandre Gil, and Marco Antonio Pinheiro Machado, *bacia de santos*, Boletim de Geociencias da PETROBRAS **15** (2007), no. 2, 531–549. ↑12, 20, 23, 27, 28, 29, 30

[MMK03] David JC MacKay and David JC Mac Kay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003. ↑40

[MS90] Susan LM Miller and Robert R Stewart, *Effects of lithology, porosity and shaliness on p-and s-wave velocities from sonic logs*, Canadian Journal of Exploration Geophysics **26** (1990), no. 1-2, 94–103. ↑49

[Nat74] JF Nations, *Lithology and porosity from acoustic shear and compressional wave transit time relationships*, Spwla 15th annual logging symposium, 1974. ↑49

[NG17] M Nogueira and RV Gaier, *Petrobras vê custo no pré-sal abaixo de us 7/barril e atrasos em plataformas*, Acesso em, 2017. ↑27

[NH10] Vinod Nair and Geoffrey E Hinton, *Rectified linear units improve restricted boltzmann machines*, Icml, 2010. ↑42

[Nic09] Gary Nichols, *Sedimentology and stratigraphy*, John Wiley & Sons, 2009. ↑19

[PF97] Colin C Potter and Darren S Foltinek, *Formation elastic parameters by deriving s-wave velocity logs*, CREWES report **9** (1997), 10–23. ↑49

[Pic63] George R Pickett, *Acoustic character logs and their applications in formation evaluation*, Journal of Petroleum technology **15** (1963), no. 06, 659–667. ↑49

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., *Scikit-learn: Machine learning in python*, the Journal of machine Learning research **12** (2011), 2825–2830. ↑39

[RdC08] André Luis Debiaso Rossi and André CPLF de Carvalho, *Bio-inspired optimization techniques for svm parameter tuning*, 2008 10th brazilian symposium on neural networks, 2008, pp. 57–62. ↑40

[RHG+80] LL Raymer, ER Hunt, John S Gardner, et al., *An improved sonic transit time-to-porosity transform*, Spwla 21st annual logging symposium, 1980. ↑49

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, *Learning representations by back-propagating errors*, nature **323** (1986), no. 6088, 533–536. ↑42

[RKK19] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar, *On the convergence of adam and beyond*, arXiv preprint arXiv:1904.09237 (2019). ↑43

[RM17] Sebastian Raschka and Vahid Mirjalili, *Python machine learning: Machine learning and deep learning with python*, Scikit-Learn, and TensorFlow. Second edition ed (2017). ↑34

[RST12] Claudio Riccomini, Lucy Gomes Sant, and Colombo Celso Gaeta Tassinari, *Pré-sal: geologia e exploração*, Revista Usp **95** (2012), 33–42. ↑12, 31

[Rus44] William Low Russell, *The total gamma ray activity of sedimentary rocks as indicated by geiger counter determinations*, Geophysics **9** (1944), no. 2, 180–216. ↑45

[RWS+14] Qiquan Ran, Yongjun Wang, Yuanhui Sun, Lin Yan, and Min Tong, *Volcanic gas reservoir characterization*, Elsevier, 2014. ↑44, 46

[SA80] O Serra and H Abbott, *The contribution of logging data to sedimentology and stratigraphy.(spe paper 9270.) paper presented at the 55th annual fall technical conference and exhibition of the society of petroleum engineers of aime*, Dallas, TX (1980), 21–24. ↑19

[SAKA12] Ebrahim Sfidari, Abdolhossein Amini, Ali Kadkhodaie, and Bahman Ahmadi, *Electrofacies clustering and a hybrid intelligent based method for porosity and permeability prediction in the south pars gas field, persian gulf*, Geopersia **2** (2012), no. 2, 11–23. ↑18

[Sau16] Ildo L Sauer, *O pré-sal e a geopolítica e hegemonia do petróleo face às mudanças climáticas e à transição energética*, Recursos Minerais do Brasil (2016). ↑20, 26, 27

[SBNR19] FGM Silva, Carlos Francisco Beneduzi, Gabriel Feres Nassau, and TB Rossi, *Using sonic log to estimate porosity and permeability in carbonates*, 16th international congress of the sbgf held in rio de janeiro, brazil, 2019, pp. 19–22. ↑48

[SCS19] Leonardo Silveira de Souza and Geraldo Norberto Chaves-Sgarbi, *Bacia de santos no brasil: geologia, exploração e produção de petróleo e gás natural*, Boletín de Geología **41** (2019), no. 1, 175–195. ↑26

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, *Dropout: a simple way to prevent neural networks from overfitting*, The journal of machine learning research **15** (2014), no. 1, 1929–1958. ↑43

[SI16] Abdus Satter and Ghulam M Iqbal, *Reservoir rock properties*, Reservoir Engineering (2016), 29–79. ↑47

[SJ93] Patrick L Scholes and Dave Johnston, *Coalbed methane applications of wireline logs: Chapter 13* (1993). ↑47

[SKP97] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal, *Introduction to multi-layer feed-forward neural networks*, Chemometrics and intelligent laboratory systems **39** (1997), no. 1, 43–62. ↑42

[SPBW16] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills, *Svm parameter optimization using grid search and genetic algorithm to improve classification performance*, Telkomnika **14** (2016), no. 4, 1502. ↑40

[Spe91] Donald F Specht, *A general regression neural network*, IEEE transactions on neural networks **2** (1991), no. 6, 568–576. ↑18

[SR11] Uma Shankar and Michael Riedel, *Gas hydrate saturation in the krishna–godavari basin from p-wave velocity and electrical resistivity logs*, Marine and Petroleum Geology **28** (2011), no. 10, 1768–1778. ↑48

[Sri13] Nitish Srivastava, *Improving neural networks with dropout*, University of Toronto **182** (2013), no. 566, 7. ↑43

[SS04] Alex J Smola and Bernhard Schölkopf, *A tutorial on support vector regression*, Statistics and computing **14** (2004), no. 3, 199–222. ↑38

[SS16] LS Souza and GNC Sgarbi, *Bacia de santos: de promissora a principal bacia produtora de hidrocarbonetos do brasil*, Xlviii congresso brasileiro de geologia, 2016. ↑26, 27

[STS16a] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma, *A review of supervised machine learning algorithms*, 2016 3rd international conference on computing for sustainable global development (indiacom), 2016, pp. 1310–1315. ↑33, 36

[STS16b] ———, *A review of supervised machine learning algorithms*, 2016 3rd international conference on computing for sustainable global development (indiacom), 2016, pp. 1310–1315. ↑37

[SV99] Johan AK Suykens and Joos Vandewalle, *Least squares support vector machine classifiers*, Neural processing letters **9** (1999), no. 3, 293–300. ↑37

[TB20] Andre Torres and Alessandro Batezelli, *Applying supervised machine learning model to classify electrofacies in a brazilian pre-salt wellbore* (2020). ↑18, 19

[TFMA08] Antonio Thomaz Filho, Ana Maria Pimentel Mizusaki, and Luzia Antonioli, *Magmatismo nas bacias sedimentares brasileiras e sua influência na geologia do petróleo*, Revista Brasileira de Geociências **38** (2008), no. 2 suppl, 128–137. ↑27

[Tit12] Jay Tittman, *Geophysical well logging: excerpted from methods of experimental physics*, Elsevier, 2012. ↑23, 44, 47

[Vap13] Vladimir Vapnik, *The nature of statistical learning theory*, Springer science & business media, 2013. ↑18

[Vap95] Vladimir N Vapnik, *The nature of statistical learning*, Theory (1995). ↑37, 38

[Vap98] Vladimir Vapnik, *The support vector method of function estimation*, Nonlinear modeling, 1998, pp. 55–85. ↑37, 38

[VPS18] HP Vinutha, B Poornima, and BM Sagar, *Detection of outliers using interquartile range technique from intrusion dataset*, Information and decision sciences, 2018, pp. 511–518. ↑55

[WBH+02] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu, *A comparative study of rnn for outlier detection in data mining*, 2002 ieee international conference on data mining, 2002. proceedings., 2002, pp. 709–712. ↑58

[WCRS01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl, *Constrained k-means clustering with background knowledge*, Icml, 1, 577-584, 2001. ↑34

[WDW+15] Ru-yue WANG, Wen-long DING, Zhe Wang, Ang Li, Jian-hua HE, and Shuai Yin, *Progress of geophysical well logging in shale gas reservoir evaluation*, Progress in Geophysics **30** (2015), no. 1, 228–241. ↑23

[WGG56] Malcolm Robert Jesse Wyllie, Alvin Ray Gregory, and Louis Wright Gardner, *Elastic wave velocities in heterogeneous and porous media*, Geophysics **21** (1956), no. 1, 41–70. ↑49

[WJW90] ML Wiggins and HC Juvkam-Wold, *Simplified equations for planning directional and horizontal wells*, Spe eastern regional meeting, 1990. ↑24

[WLZ+18] Meiqi Wang, Siyuan Lu, Danyang Zhu, Jun Lin, and Zhongfeng Wang, *A high-speed and low-complexity architecture for softmax function in deep learning*, 2018 ieee asia pacific conference on circuits and systems (apccas), 2018, pp. 223–226. ↑42

[XTJ17] Weiming Xiang, Hoang-Dung Tran, and Taylor T Johnson, *Reachable set computation and safety verification for neural networks with relu activations*, arXiv preprint arXiv:1712.08163 (2017). ↑18

[Yan07] Yuhong Yang, *Consistency of cross validation for comparing regression procedures*, Annals of Statistics **35** (2007), no. 6, 2450–2473. ↑64

# Appendix A

# K-means algorithm approach for automate the electrofacies classification: An exploratory study applied in Brazilian Pre-Salt, Santos Basin

# K-means algorithm approach for automate the electrofacies classification: An exploratory study applied in Brazilian Pre-Salt, Santos Basin.

*André L P Torres*, Alessandro Batezelli, State University of Campinas*

**Summary**

Organizing data sets in cluster has been one of the greatest challenge in the modern time, mainly because in the last years, the amount of data acquired has increased a lot. In petroleum geology, analyzing a large volume of well log data in order to extract reservoir properties by manual approaches is difficult and time consuming. Thus, methods and algorithms that offer to classify these clusters has become great allies to better understand the big amount of data and information. The K-means is an unsupervised machine learning algorithm that through of the sum of square errors proposes to divide a data set in K clusters. In the petroleum reservoir characterization, the electrofacies are a theme quite discussed, mainly because they reflected the properties of the rock, fluids and pores. Thus, geological models can be build and, consequently, they can auxiliary to minimize the uncertainties in exploration. The work used the K-means algorithm to classify different electrofacies in a data set which were selected eight features. According to Elbow Method, the ideal number of clusters for the data set was two, thus the data set was classified in two electrofacies. The K-means algorithm revealed to be a robust method for grouping electrofacies and a promising method capable of automate the processing of electrofacies classification. The research shows new ways to classify electrofacies and try to open new possibilities for machine learning algorithms in the geoscience world.

## Introduction

Electrofacies have been used widely in the petroleum prospecting and reservoir characterization as a tried to distinguish different beds in a petroleum field as well as in the correlation with the lithofacies. The main difference between electrofacies and lithofacies is that the first represent a response of a set log records of the rock, pores and fluids, while the second is define as a rock unit defined by distinct lithological features, including composition, granulometry and sedimentary structures. Therefore is important to understand that electrofacies and lithofacies will not always mach each other. Moreover, petrophysic is another way to determine reservoir properties,but this method is not useful for sedimentary studies, separation and zoning of reservoirs,and geological properties (Kiaei et al., 2015). Traditionally electrofacies has been identified manually with the aid of graphical techniques like crossplotting from wire-line logs and thereby correlating their behavior to cores. Most recently several machine learning algorithm have been introduced to try automate the task of facies identification (Kumar and Kishore, 2006).

Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning, and cluster analysis is the formal study of methods and algorithms for grouping, objects according to measured or perceived intrinsic characteristics or similarity (Jain, 2010). The K-means clustering (MacQueen et al., 1967) is a method used to grouping into k groups a data set, and it has a wide applicability in many knowledge areas.

The work proposes to use the K-mean method to analyze the electrofacies of the geophysical dataset of the Gato do Mato Oil Field, Santos Basin, in order to identify different electrofacies related to carbonate reservoir rocks. The choice to specifically study area is due to its importance in become the Brazilian pre salt more understandable and access. The Santos Basin is localized in the southeast of Brazil, with an area approximate of 350.000 $km^2$ and the sediment thickness in some areas is higher than 10 km (Chang et al., 2008). The figure 1 presents the Pre-salt distribution, as well as the study area.

The figure 1 shows the localization of the study area.

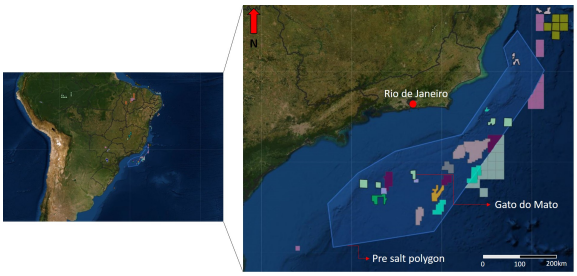



Figure 1: Localization of the Gato do Mato field in the brazilian Pre-salt polygon.

## Theory and Method

The research worked with the K-means algorithm and Elbow method together to classify and partition the data set. How is explained forward the K-means algorithm needs that the external user input the cluster number(K) that the data set will be partition. One way to find the correct cluster number is to know very well the data set(specialist), the other way, the automate way, is applying the Elbow Method. In the present work was used the second way to find the correct cluster number.

1. **Elbow Method**

   Elbow method is a method which looks at the Sum Square error percentage of variance explained as a function of the number of clusters (Bholowalia and Arvind, 2014). The Elbow method is expressed by

$$SSE = \sum_{K=1}^{K} \sum_{Xi \in Sk} \| Xi - Ck \|_2^2 \qquad (1)$$

   Where SSE is the sum of the average Euclidean Distance of each point against the centroid (Marutho et al., 2018).The letter K is the number of clusters, Xi is the

data present in each cluster, Ck is the K-th cluster and Sk is the set of points inside the Ck cluster.

The Sum Square error explained by the cluster is plotted against the number of clusters. The first cluster will add much information, but at some point the marginal gain will drop dramatically and gives an angle in the graph (Bholowalia and Arvind, 2014). When the increase of the cluster number does not varies considerably the SSE, then the best K value was found.

Most of clustering algorithms are designed only to investigate the inherited grouping or partition of data objects according to a known number of clusters. Thus, identifying the number of clusters is an important task for any clustering problem (Kodinariya and Makwana, 2013).

2. **K-means Algorithms**

K-means clustering is a method commonly used to automatically partition a data set into k groups (Wagstaff et al., 2001).The data clustering, also known as cluster analysis, try to discover the natural grouping(s) of a set of patterns, points, or objects (Jain, 2010). The aim of cluster analysis is to classify a data set into groups that are internally homogeneous and externally isolated on the basis of a measure of similarity or dissimilarity between groups (Kiaei et al., 2015).

Several clusters algorithms have been proposed to try classify different data set, but due to its simplicity, the K-means algorithm have been the most commonly used in the literature. Given a set of $n$ data points in real $d-dimensional$ space, $\mathbf{R}^d$, and an integer $k$, the problem is to determine a set of $k$ points in $\mathbf{R}^d$, called centers, so as to minimize the mean squared distance from each data point to its nearest center (Kanungo et al., 2002).

Let $X = \{x_i\}, i = 1,...,n$ be the set of n d-dimensional points to be clustered into a set of K clusters, $C = \{c_k,\ K = 1,...,k\}$. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized (Jain, 2010). The K-means works to minimize the sum of the squared error (Eq.1). Minimizing this objective function is known to be an NP-hard problem (even for K = 2) (Drineas et al., 1999). K-means is better explained in the steps (I) to (V) and by the work flowchart illustrated in the Fig.2

(I) Assigning the centroids;

(II) The distance between each point to the centroid is calculated. N-points and K-centroids;

$$N \times K = number\ of\ distances\ calculated$$

(III) Each point is putting in the class according to the centroid distance. The point is embody by the nearest centroid and will belong to the class represented by centroid;

(IV) New centroids are calculated for each class and the value of centroid coordinate are refined. For each class that has more than one point, the new centroid coordinate is calculated using coordinate average of all points belongs to the class;

(V) The algorithm repeat the third and fourth step repeatedly until the convergence. When in the loop n the centroid coordinate doesn't change in relation to the previous loop (n-1), then the process finish and the centroid coordinate is found.

The figure 2 shows the K-means work flowchart.

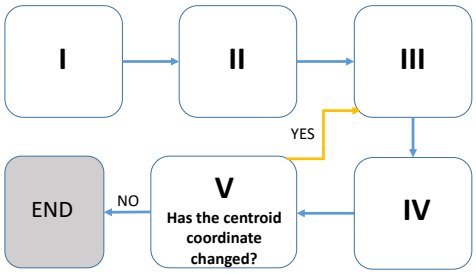



Figure 2: Work flowchart of the K-Means algorithm.

In the K-means algorithm the users need to provide the number of classes that fulfill their wishes, but the wrong choose of the number of clusters will result in K-means clustering algorithm with high erros and poor's cluster results. Thus, the Elbow Method can be an important and complementary method to help in the choose of the correct number of clusters in a dataset.

**Examples**

The data set from well Shel-23 was provide by the National Agency of Petroleum (ANP) from Brazil. In the research was used Petrel® software and Jupyter® notebook to work the data set Fig.3, the part of data science and machine learning algorithms was implemented in python language.

The figure 3 shows the data set statistic.

| | Depth | gamma | density | neutron | porosity | Resistivity | Sw | P-vel. | S-vel. |
|---|---|---|---|---|---|---|---|---|---|
| count | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 | 1814.000000 |
| mean | 5253.914201 | 19.781178 | 3.235072 | 0.090143 | 0.046361 | 771.014036 | 0.585359 | 4999.955375 | 2694.346819 |
| std | 79.827272 | 11.511092 | 1.004592 | 0.064277 | 0.052591 | 644.051385 | 0.395095 | 490.270627 | 295.021824 |
| min | 5115.760000 | 1.562000 | 1.988900 | -0.017300 | -0.076400 | -62.907000 | 0.001100 | 4017.930000 | 2111.060000 |
| 25% | 5184.837500 | 8.178575 | 2.471900 | 0.040725 | -0.000575 | 84.489500 | 0.173600 | 4543.565000 | 2451.160000 |
| 50% | 5253.915000 | 20.551700 | 2.892250 | 0.098500 | 0.055600 | 799.922500 | 0.535700 | 4936.585000 | 2659.600000 |
| 75% | 5322.992500 | 27.683550 | 4.626475 | 0.138775 | 0.087475 | 1351.196000 | 1.000000 | 5446.477500 | 2890.647500 |
| max | 5392.060000 | 86.376800 | 4.628800 | 0.238700 | 0.162500 | 1950.000000 | 1.000000 | 6101.350000 | 3438.330000 |

Figure 3: Data set statistic from Shel-23.

The K-means algorithm was applied in eight logs profile to classify the electrofacies using an unsupervised machine learning algorithm (K-means). Some of these logs have been used in the main electrofacies works to classify the differents electrofacies (Kumar and Kishore, 2006).

There are three main properties that the logs can measure in the formation, they are, rock, pores and fluids, in the work was used, Gamma ray to verify the rock composition, the $V_p$ and

$V_s$ to assimilate elastic properties, the density, effective porosity, SPHI and neutron to measure the porous level and the resistivity and water saturation to indicate the fluids properties. The first point that must be taken attention was to normalize or standardize all data set features, this ensured that all features had the same weight in the algorithm. The second caution was to find the correct number of clusters, for auxiliary in this process was used the Elbow Method, previously explained, but instead to use SSE, in the current work was used the Inertia, the idea the same, good clustering is having a small value of Inertia, and small number of clusters. The figure 4 shows the Elbow Method applied in the data set.

The figure 4 shows the Elbow Method.

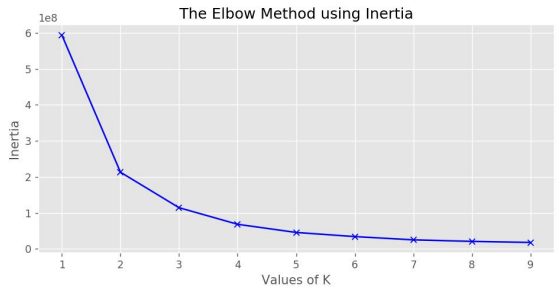



Figure 4: The Elbow Method applied in the data set.

According to the figure 4 the ideal number K of cluster is 2. Thus, with two clusters the K-means algorithm was running. The figure 5 represent the eight log profiles used in the K-means, and the last profile, the electrofacies, is the result of the K-means algorithm with two clusters in the data set (fig 3). The K-means algorithm together with Elbow Method shows to be an robust combination in clustering problens. Analysing the figure 5 the main features used by the algorithm seems be gamma ray, $V_p$ and $V_s$, the electrofacies (V) or yellow electrofacies was related with lowest $V_p$ and $V_s$ values, and high gamma ray values, the other features did not have a clear relation. How the data set was classify in two clusters, the electrofacies (VI) or pink electrofacies was associated with high values of $V_p$ and $V_s$, and lowest values of gamma ray.

**Conclusions**

The firsts results obtained with the K-means application can be considered promising once time that the algorithm showed to be capable of precisely differentiate the data set. In other words, the K-means algorithm shows a high potential for automate the process of classify electrofacies, this automation if well calibrated can save time and resources of petroleum companies in large fields with a lot well logs informations. Another advantage of the K-means algorithm into classify the electrofacies is that the algorithm can to analyse several features, and thus, to find patterns difficult to be find by human eyes, besides that after clustering log data, the log response can be generalized to the overall reservoir and thus log data that represents a few inches of the surrounding well can be interpolated by the wells in the field and to model all reservoir.

In the current work noted that for the human eyes only a few features could be correlated with the electrofacies, this is a point that must to be work in the next steps for a better understanding. The current research approached themes of high importance in the modern scenario, the union between well logs data set and machine learning algorithm have showed a success combination. Therefore, the work contribute for the geoscience and machine learning literatures as a practical case of machine learning algorithms automating an important process in the petroleum geology.

**Acknowledgments**

I wish to thank Alessandro Batezelli for help in my master and in current work. I wish to thank the Brazilian National Agency of Petroleum for providing the data set that was used in this work. I wish to thank the Faculty of Mechanical Engineering(FEM) and Petroleum Engineering Department (DEP) to contribute with the all physical structure, buildings and laboratories, that enabled the project to be carried out.

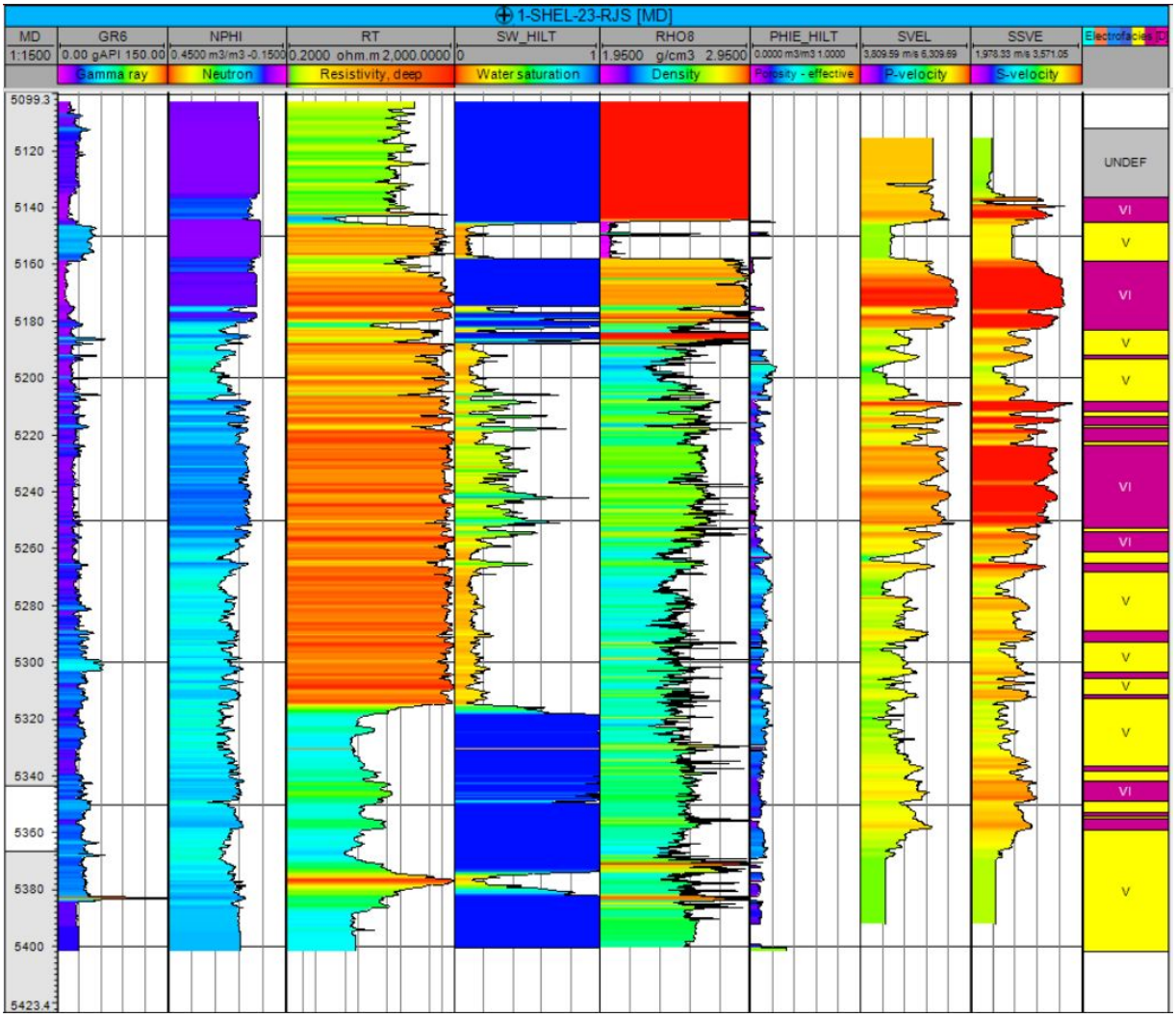



Figure 5: Well logs profile with the electrofacies classification by K-means algorithm.

**REFERENCES**

Bholowalia, P., and K. Arvind, 2014, Ebk-means: A clustering technique based on elbow method and k-means in wsn: International Journal of Computer Applications 105(9).

Chang, H. K., M. L. Assine, F. S. Corrêa, J. S. Tinen, A. C. Vidal, and L. Koike, 2008, Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na bacia de santos: Revista Brasileira de Geociências, **38**, 29–46.

Drineas, P., A. M. Frieze, R. Kannan, S. Vempala, and V. Vinay, 1999, Clustering in large graphs and matrices.: SODA, Citeseer, 99, 291–299.

Jain, A. K., 2010, Data clustering: 50 years beyond k-means: Pattern recognition letters, **31**, 651–666.

Kanungo, T., D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, 2002, An efficient k-means clustering algorithm: Analysis and implementation: IEEE transactions on pattern analysis and machine intelligence, **24**, 881–892.

Kiaei, H., Y. Sharghi, A. K. Ilkhchi, and M. Naderi, 2015, 3d modeling of reservoir electrofacies using integration clustering and geostatistic method in central field of persian gulf: Journal of Petroleum Science and Engineering, **135**, 152–160.

Kodinariya, T. M., and P. R. Makwana, 2013, Review on determining number of cluster in k-means clustering: International Journal, 1(6), 90-95.

Kumar, B., and M. Kishore, 2006, Electrofacies classification—a critical approach: 6th international conference and exposition on petroleum geophysics, Kolkata, India, 822–825.

MacQueen, J., et al., 1967, Some methods for classification and analysis of multivariate observations: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, 281–297.

Marutho, D., S. H. Handaka, E. Wijaya, et al., 2018, The determination of cluster number at k-mean using elbow method and purity evaluation on headline news: 2018 International Seminar on Application for Technology of Information and Communication, IEEE. 533–538.

Wagstaff, K., C. Cardie, S. Rogers, S. Schrödl, et al., 2001, Constrained k-means clustering with background knowledge: Icml, 1, 577-584.

# Appendix B

# Applying supervised machine learning model to classify electrofacies in a Brazilian Pre-salt wellbore

Rio Oil & Gas Expo and Conference 2020

ISSN 2525-7579

Conference Proceedings homepage: https://biblioteca.ibp.org.br/riooilegas/en/

*Technical Paper*

# Applying supervised machine learning model to classify electrofacies in a Brazilian Pre-salt wellbore

**Andre Luiz Pontara Torres** [1]
**Alessandro Batezelli** [2].

1. UNICAMP, FACULDADE DE ENGENHARIA MECâNICA, DEPARTAMENTO DE ENERGIA(DE)/DEPARTAMENTO DE ENGENHARIA DE PETRóLEO(DEP). CAMPINAS - SP - BRASIL, a090388@dac.unicamp.br
2. UNICAMP, INSTITUTO DE GEOCIêNCIAS(IGE), DEPARTAMENTO DE GEOLOGIA E RECURSOS NATURAIS(DGRN). CAMPINAS - SP - BRASIL, batezeli@unicamp.br

**Abstract**

Machine learning algorithms have been more and more present in the current days in all knowledge areas, mainly when the objective is treatment of data. In the petroleum geology these algorithms have been used frequently in the seismic and well logs data. Applying the Machine Learning algorithms, the work has proposed to classify different electrofacies from well logs data set. The data set is from pre-salt, located in Santos Basin–Brazil. The classification was made using the Support vector machine (SVM) and in the initial model, the research has reached a hit rate of 86%. The first results can be considered promising once there are many parameters that can be refining in the next steps, what can implying in a possible increase of the hit rate.

**Keywords:** Support vector machine. Petroleum Geology. Electrofacies. Pre-salt carbonates

**Received:** February 28, 2020 | **Accepted:** Jun 06, 2020 | **Available online:** Dec 01, 2020
**Article Code:** 014
**Cite as:** Rio Oil & Gas Expo and Conference, Rio de Janeiro, RJ, Brazil, 2020 (20)
**DOI:** https://doi.org/10.48072/2525-7579.rog.2020.014

## 1. Introduction

In recent years Machine Learning algorithms have been becoming more and more present in every area of knowledge, and in the petroleum geology is not different. A lot of works trying to apply Machine Learning algorithms can be found in the recent literature (Kuroda et at., 2012; Kadkhodaie & Ahmadi, 2012), however, due to being a new field of study, there are many challenges to be overcome. Machine Learning algorithms are responsible for data set treatment for the purpose of classification and prediction. Among the Machine Learning algorithms there are two major groups, the supervised Machine Learning algorithms (Kotsiantis, 2007) and the unsupervised Machine Learning algorithms (Jain, 2010).

In petroleum geology an indirect way to identify the rock properties in a subsurface is through the well-logs, that are responsible for indirectly measuring the rock properties in the subsurface. These records are able to identify the, porosity, density, resistivity and radioactivity in the reservoir. In order to organize the well-log data set from a graphic and statistical analysis the geologists try to recognize patterns to distinguish different zones in a subsurface. Those zones identified by the Geologists based on from analysis of well-log data set are called electrofacies. But people are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems (Kiaei et al., 2015), improving the efficiency of systems and the designs of machines (Kotsiantis, 2007).

Estimates in Santos Basin suggest that the potencial volume of oil reserves is higher than 100 billion barrels (Sauer, 2016), would position Brazilas having the fifth biggest world reserves. The stratigraphic section studied in the Santos Basin is localized under the evaporitic unit formed during the pos rift phase in the Aptian Stage. The main stratigraphic units that compose the study area are Itapema and Piçarras Formations. The Itapema is located immediately bellow the evaporitic section and was formed between the Late Barremian to Early Aptian. Based on the paleogeographic distribution, in the distal portions, the Itapema Formation were formed by marine incursions that were responsible to deposited dark shales and carbonate rocks (Arai, 2014), while the proximal portions are constituted by conglomerates and sandstones deposited by alluvial fans (Moreira et al., 2007). The Piçarras Formation correspond to those alluvial fans sediments, composed by conglomerates and lithic sandstones deposited during the Barremian Stage. Volcanic rocks of the Camboriú Formation(Upper Necomian) constitute the basement of the basin (Moreira et al., 2007).

Electrofacies is a term used since the 80's (Serra & Abbott, 1980) to describe a set of logs that have some similar outputs. Electrofacies represent a unique set of log responses, which characterize the physical properties of the rocks and fluids contained in the volume investigated by the logging tools (Euzen & Power 2014). Traditionally electrofacies have been identified manually with the aid of graphical techniques like crossplotting from wire-line logs (Kumar & Kishore, 2006).

In this work is purposed the application of Machine Learning algorithms in well-logs, from Santos Basin, to make the classification of electrofacies from log responses. The method used to make the classification and prediction was the Support Vector Machine, a complex supervised algorithm able to provide high accuracy in problems of prediction and classification.

The current work contributed to the discussion about the importance and value of Machine Learning algorithms in electrofacies classification and prediction, and overcoming the challenges involving this process could result in more precision and agility in electrofacies classification. Thus, it is crucial for the results and discussion about this theme to become more visible in the petroleum literature.

## 2. Development

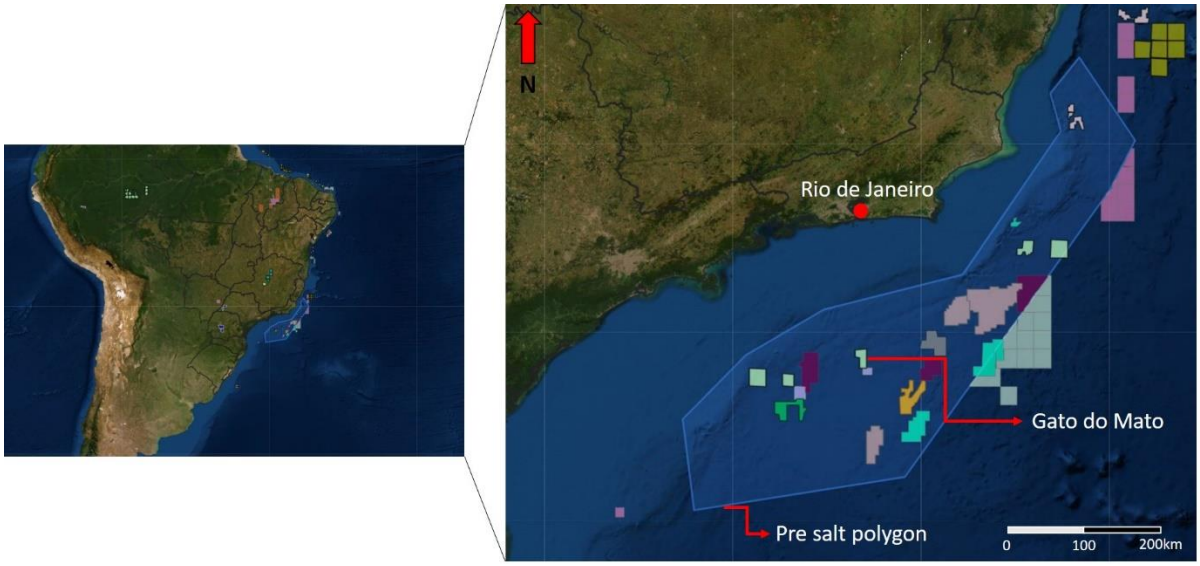
Due to the large amount of data collected by the petroleum industry, the use of robust algorithms in data processing and analysis has become essential to guarantee efficiency, speed and accuracy. In this context Artificial Intelligence algorithms such as Machine Learning and Deep Learning have been applied by the Petroleum Industry to prediction and classification. Some specific works to the O&G area have been done and different application has emerged, Kumar and Kishore (2006) show a way to identify the lithological and depositional facies from wireline logs using an approach based on Feed Forward Neural Network and Clustering, Kuroda, Vidal, Leite and Drummond (2012) also proposing an electrofacies characterization using artificial neural network from petrophysical data, such as, neutron, porosity, gamma ray, density and sonic profiles. With a data set of well logs from Amazonia, Oliveira Júnior (2014) analyzed the performance of five machine learning algorithms to classify electrofacies, the results shown that SVM was the better classifier in two of three well log. Sfidari, Amini, Kadkhodaie and Ahmadi (2012) characterized reservoir properties with a geological and petrophysical integration, the properties porosity and permeability were predicted by linear functions while the identification and extration of electrofacies groups were made with an unsupervised neural network (SOM). Lee, Khanghoria and Datta-Gupta (2002) developmented a methodology to classify electrofacies based on three unsupervised learning algorithms, they are: principal components analysis (PCA), Cluster Analysis and Discriminat analysis. Another important work produced by, Kiaei, Sharghi, Ilkhchi and Naderi (2015) also shown how the application of Artificial intelligence algorithms can help in the petroleum industry, in a 3D electrofacies modeling of a reservoir , machine learning algorithms were used to classify electrofacies and save time.

Following the same reasoning of the works cited before, the current work dealt with well-logs data set and applying a supervised Machine Learning algorithm proposed the classification and prediction of electrofacies.

The well-logs data set was provided by the ANP, they are located in the Gato do Mato oilfield area. Based on the well-logs, were obtained information about rock (Gamma-Ray, P-velocity and V-velocity), fluids (Resistivity, Water Saturation) and porous (Neutron, sphi and density). Those data were analysed to elaborate the classification and prediction of the electrofacies (figure 4).

The Santos Basin is localized on the southeast margin of Brazil, with an area is approximately 350.000km², and the sediment thickness in some areas is higher than 10 km (Chang et al., 2008) (figure 1). The study area in Gato do Mato oilfield is located in the southwest of this basin. The study used information over two well-logs from Pre Salt reservoir, with more than 5.000 meters of depth. The interval of interest has up to 400 meter and presents geophysic and lithological characteristics of the carbonate rocks, used to the classification and prediction by the machine learnig methods.

**Figure 1 -** Location of pre salt polygon and study area.




**Source:** Produced by the author.

The classification and prediction of the electrofacies was made using the SVMs algorithm and achieved expressived results. To reach the results was necessary work the data set in order to become more reliable and in correct format to be read by the algorithm.

## 2.1. Theoric fundamentation

The Support Vector Machine is an algorithm to predict and classify that was developed by Vladimir Vapnik in 1960s, and in the nineties the algorithm was highlighted to solving pattern recognition problems through of the Vapnik's works (Vapnik, 1995; Vapknik, 1998a; Vapnik, 1998b).

The SVMs method maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space. The algorithm works with a concept of maximizing the minimum distance from hyperplane to the nearest sample point (Singh et al., 2016). This basically involves solving a quadratic programming problem (Suykens & Vandewalle, 1999).

The technique of SVMs was first developed for the restricted case of separating training data without errors, later was enhanced the case of separating data with erros. Thus, there are two historic Support Vector Machine algorithm, (i) the hard margin Support Vector Machine and (ii) the soft margin Support Vector Machine that will allow for an analytic treatment of learning with errors on the training set (Cortes & Vapnik, 1995). In both algorithms the elements, Hyperplane, Margin and Support vectors are essential.

The hyperplane is understood as a plane that separate the data, the margin is the distance of the hyperplane and the support vectors and the support vectors are the closest points of the hyperplane. The optimal hyperplane is defined as the linear decision function with maximal margin between the vectors of the two classes (Cortes & Vapnik, 1995). According to Cortes and Vapnik (1995) to find the optimal hyperplane in the hard margin case the algorithm considers the set of labeled training patterns:

$$(y_1, x_1), \dots, (y_l, x_l), \qquad y_i \in \{-1, 1\} \qquad (1)$$

The dataset is linearly separable if there is a vector W and a scalar b such that

$$y_i(W \cdot x_i + \mathrm{b}) \geq 1 \qquad \mathrm{i} = 1, \ldots, \mathrm{L}. \qquad (2)$$

are valid for all elements of the training set. The optimal hyperplane is given by

$$W_0 \cdot x + b_0 = 0 \qquad (3)$$

The maximal distance between the projections of the training vectors of two different classes under the constraints (2) will ensure an optimal hyperplane (3)

$$\varphi = W \cdot W \qquad (4)$$

for this the equation (4) has to be minimized. To do this a standard optimization technique is used and a Lagrangian is constructed

$$L(W, b, \Lambda) = \frac{1}{2} W \cdot W - \sum_{i=1}^{L} \alpha_{\mathrm{i}}[y_i(x_i \cdot W + b) - 1] \qquad (5)$$

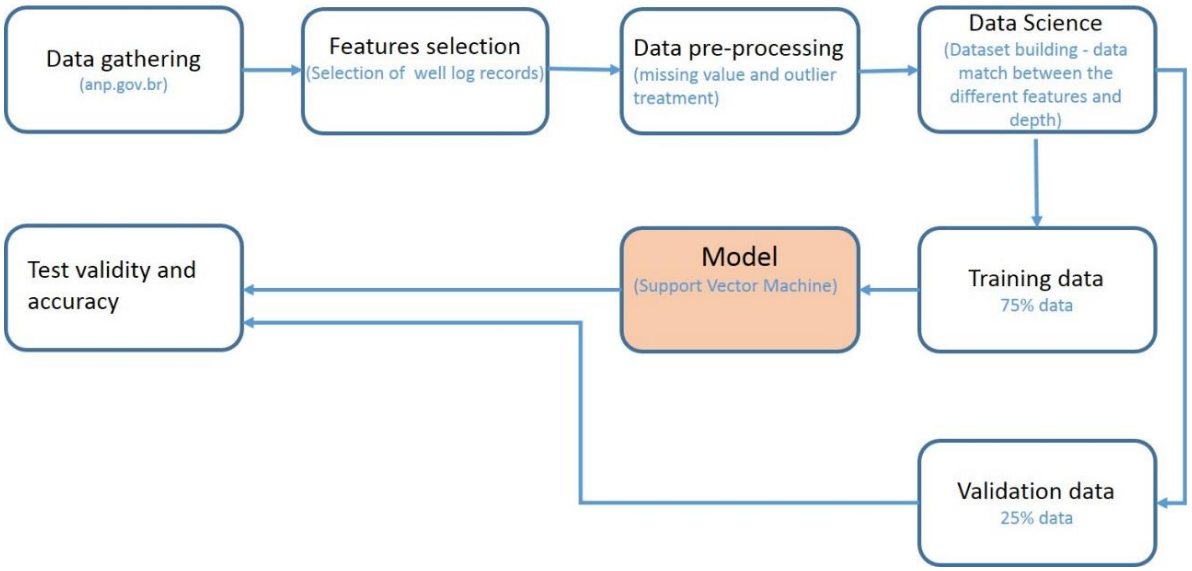
where $\Lambda^T = (\alpha_1, \ldots \alpha_l)$ is the vector of non-negative Lagrange multipliers corresponding to the constraints (2). To find the hyperplane in the soft margin case the reasoning is similar and can be found in the article of Cortes and Vapnik (1995).

The SVMs algorithm deals very well with linear data sets, in cases where a straight line is sufficient to separate a data set, but for problems with non linear data sets isn't possible to find a straight line to classify the data set, in this case the Kernel Trick (Boser, Guyon & Vapnik, 1992) is used. The basic idea is that if a data set is inseparable in the current dimensions, the kernel trick will carry the input data set to a higher dimensional space and by choosing an adequate dimension, the data set points become linearly separable or mostly linearly separable in the high-dimensional space (Amari & Wu, 1999). Thus, the SVMs with the Kernel trick currently has been ensured the solving of linear and non linear problems efficiently.

## 2.2. Methodology

The processing of data was made in three different environments, they are Petrel software, Excel and Jupyter Notebook. In the first step was created the data set, selecting profiles and organizing well-logs data set according to the depth of the sampling. The construction and treatment of the data set was made with the auxiliary of some macros in Excel and the creation and application of some data science algorithms in the Jupyter Notebook. The work fluxogram can be seen in the figure 2.

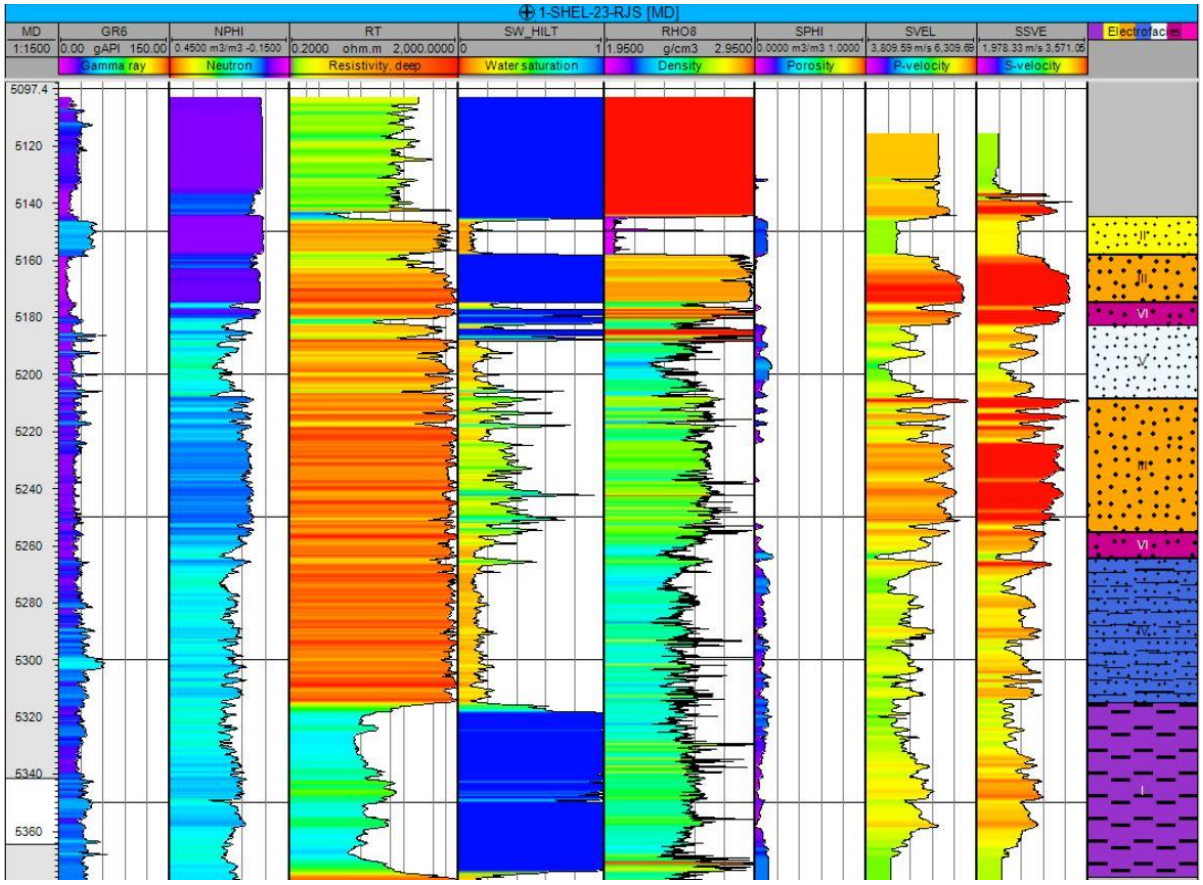
**Figure 2 -** Methodology applied in the work.




**Source:** Produced by the author.

With the data set constructed, a manual electrofacies classification of sendimentary facies (carbonate rocks) from a well-log was made. The classification was made according to electrical characteristics records from rocks in subsurface. Petrel software was used to plot logs and, thus, to help in a better visualization and comprehension of the patterns. A manual classification (figure 3) was made in the 1-SHEL-23-RJS well-log, and was used as the reference to electrofacies classification from others well-logs. After manual electrofacies classification, a supervised Machine Learning algorithm was applied in the manual classification in order to learn how to classify others data sets. The algorithm used was the Support Vector Machine (SVM), due to the fact that the algorithm works with the input data in a higher dimensional space and, thus, getting to obtain a better classification and prediction. The manual classification (figure 3) of the 1-SHEL-23-RJS data set was used to teach the SVM how to classify others data sets. For that, the 1-SHEL-23-RJS data set, previously labeled by the manual classification step, was divided in two groups: the training data set (75%) and the test data set (25%). In the training data set, the SVM has learned to classify different electrofacies from the manual classification, while in the test data set the SVM algorithm was used to predict the electrofacies according to the input features (logs). Thus, how the whole 1-SHEL-23-RJS data set was labeled by the manual classification (figure 3), the accuracy test was made comparing the predicts electrofacies by the SVM in the test data set with the labeled electrofacies created in the manual classification. The last step was apply the trained SVM classifier to predict the electrofacies of the well-log, 1-SHEL-26-RJS (figure 4).

**Figure 3** – Well-log reference, 1-SHEL-23-RJS, with manual classification of electrofacies.




**Source:** Produced by the author.

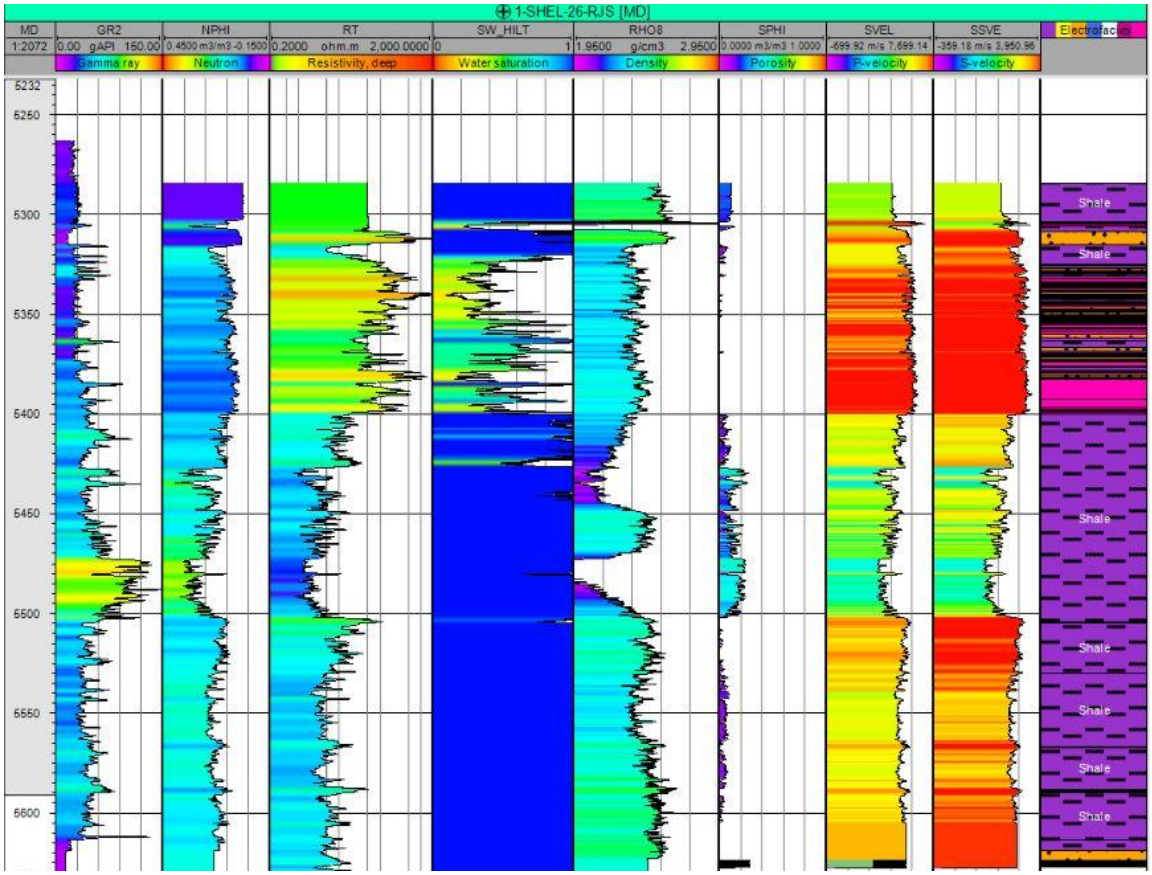
## 2.3. Results

The application of the Support Vector Machine algorithm in the data set of the 1-SHEL-23-RJS well-log, was responsable by 85.7% of hit rate, showing to be pretty promising, but it can be improved. When the algorithm was applied to predict the electrofacies of the well log, 1-SHEL-26-RJS, the results were different from the expected (Figure 4).

The first point was the low electrofacies variation in some depth (5300 – 5400m). Probably the explaining of this effect were: (i) one cause of this low generalization can be to derive of a overfitting of the algorithm in the data set. The overfitting is when the algorithm has a good performance in the training data set and a bad performance in the test data set. This happens because the algorithm learning too much from the training data set, the hypothesis (i) was verified and the hit rate in both data sets was good with the difference between them equal to 0.2%. Thus, the hypothesis (i) was disconsidered. The hypothesis (ii) could be due to the manual classification, where were found six electrofacies for, 1-SHEL-23-RJS well-log, this is a valid hypothesis because with a higher level of detail as shown in figure 3, the difference among the electrofacies can be confuse, in such a way that, the possibility of algorithm to wrong in classification is higher. The last hypothesis (iii) was that the algorithm has a high sensibility and, thus, small intervals in depth will be recognized by the algorithm

in case that occur variation in features. The second and third hypothesis seems to be the more reasonable and was better accept than the first.

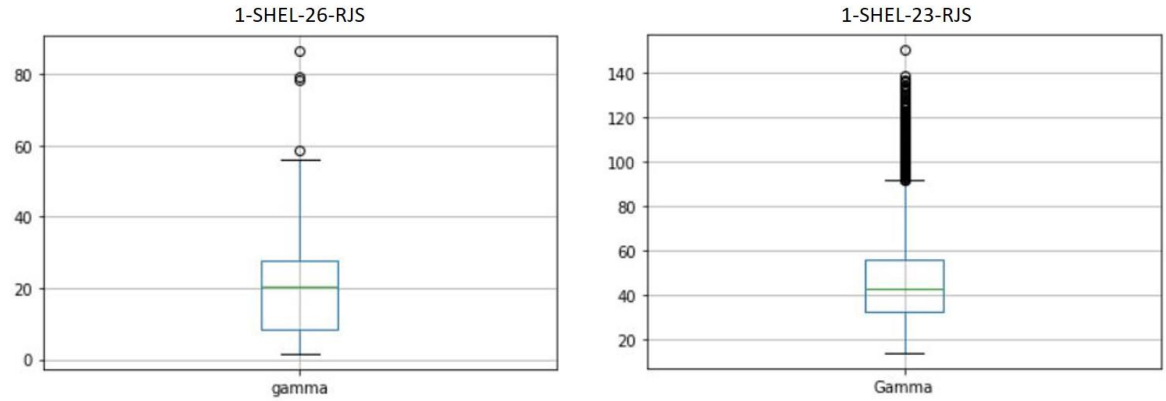
**Figure 4** – Well-log 1-SHEL-26-RJS with electrofacies predicted by SVMs.




**Source:** Produced by the author.

The second point that was noted in figure 4 was the huge amount of purple electrofacies classified. This electrofacies is characterized by high values of Gamma Ray indicating an increase in the clayiness of the formation. The huge amount of purple electrofacies happened because differently of the, 1-SHEL-23-RJS well-log, the records of Gamma Ray in the, 1-SHEL-26-RJS well-log, is high in the most part of the well. The figure 5 shows the boxplots with the Gamma-Ray distribution in both well-logs. According to figure 4 the Gamma-Ray data set of the, 1-SHEL-23-RJS, has the median above 40, the maximum value above 80 and a lot of points higher than the maximum value (outliers). In compare the Gamma-Ray data set, 1-SHEL-26-RJS, has the median equal 20, the maximum value below to 60 and few points above the maximum value (outliers). Thus, the high Gamma-Ray values found in data set of the, 1-SHEL-23-RJS well-log, explains the purple electrofacies classification.

**Figure 5 -** Gamma ray distribution represented by the boxplots.




**Source:** produced by the author.

The three electrofacies presented in the figure 4, they are, orange, pink and purple. The orange electrofacies represent sections with the lowest values of Gamma-Ray, high values of resistivity, low values of porosity, high values of water saturation and high values of P and S- velocity. The orange electrofacies represent the cleanest formation in the well. The pink electrofacies has medium value of Gamma-Ray, high values of resistivity, low values of porosity, medium values of water saturation, and the high values of P and S-velocity. The purple electrofacies has the highest values of Gamma-Ray, medium to high values of resistivity, medium values of porosity, high values of water saturation, and médium to high values of P and S-velocity.

## 3. Final Considerations

In order to recognize the patterns to distinguish different reservoir characteristics and hydrocarbons-rich zones, the present work used the Machine Learnig algorithms to propose a way to analysis the geological data set from well-logs of the Santos Basin.

The main goal of the research was to classify the electrofacies using the Support Vector Machine (SVM) and prediction the carbonatic reservoirs characteristics, based in a complex supervised algorithm able to provide high accuracy.

From a well-log reference (1-SHEL-23-RJS), was performed a manual classification of facies to work as a reference for the studies. After that, was applied the method of Support Vector Machine (SVM) to learn with the data set labeled..

Based on the reference classification, the algorithm learned to expand the electrofacies patterns to the well log (1-SHEL-26-RJS). The methodology applied in the research shows us two parameters that can be improved to make the classification process better. They are, (i) the manual classification with a lower level of detail and (ii) increasing the number of records labeled to train the model.

The results of the research shows high level of certain, and expanding the approaches by Machine Learnig to classify and predict electrofacies.

In order to become the results more confiable and decreasing the errors in the classification and prediction, is suggested that the input data set labeled be bigger and more representative, to do this a way is work with input data set labeled of more than a well-log.

How much more well-logs was used, smaller will be the errors in prediction and classification. The Support Vector Machine (SVM) shows to be an efficient algorithm for classify and predict

electrofacies in differents well-logs data, specially in the current work in which the model achieved a hit rate of 85.7%. Considering the calibration of the parameters, the current result could be better.

The high electrofacies variation in some depth (5300 – 5400m), probably was related to the detail level applied in the manual classification. With a high detail level the difference among the electrofacies can be confuse and consequently the possibility of the algorithm making mistakes in classification is higher.

The high Gamma-Ray values found in most of the well-log profile 1-SHEL-23-RJS, was the responsible to the classification as purple electrofacies.

The three electrofacies presented in the figure 4, they are, orange, pink and purple. The orange and pink electrofacies represent sections with the lowest values of Gamma-Ray, high values of resistivity, low values of porosity, high values of water saturation and high values of P and S- velocity. Those characteristics are indicating good quality of the reservoir. The purple electrofacies has the highest values of Gamma-Ray, medium to high values of resistivity, medium values of porosity, high values of water saturation, and medium to high values of P and S-velocity. Those are the main characteristics of the non-reservoir interval.

It is expected with the work that integration of manual classification and algorithms of Supervised Machine learning in well logs will improve the accuracy and increase the velocity of electrofacies classification. When the model can learn with manual classification, the SVM shows a high potential for automation of electrofacies classification process. If well calibrated, this automation can save time and resources of petroleum companies in large fields with a lot of well log information. The next steps beyond applying the considerations above, will be to make a crosscorrelation between the electrofacies and the samples described.

## 4. Acknowledgements

To the Brazilian National Agency of Petroleum (ANP) for providing the data set that was used in this work. I also wish to thanks the Faculty of Mechanical Engineering (FEM), the post graduate program and Petroleum Engineering Department (DEP) to contribute with the all physical structure, buildings and laboratories, that enabled the project to be carried out. The second author thanks to CNPq for the productivity grant (process 301200/2017-3).

Referências


Amari, S. I., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *July 1999*, (6), 783–789. https://doi.org/10.1016/S0893-6080(99)00032-5

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. (pp. 144–152). Presented at the In Proceedings of the fifth annual workshop on Computational learning theory. https://doi.org/10.1145/130385.130401

Chang, H. K., Assine, M. L., Corrêa, F. S., Tinen, J. S., Vidal, A. C., & Koike, L. (2008, June 5). Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na Bacia de Santos. *38*(2), 29–46. Retrieved from http://www.ppegeo.igc.usp.br/index.php/rbg/article/view/8161

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *September 1995*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Euzen, T., & Power, M.R. (2014). Well Log Cluster Analysis and Electrofacies Classification: A Probabilistic Approach for Integrating Log with Mineralogical Data (pp. 1–4). Presented at the CSPG/CSEG/CWLS GeoConvention 2012, Calgary, AB, Canada. Retrieved from http://www.searchanddiscovery.com/pdfz/documents/2014/41277euzen/ndx_euzen.pdf.html

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *1 June 2010*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kiaei, H., Sharghi, Y., Ilkhchi, A. K., & Naderi, M. (2015). 3D modeling of reservoir electrofacies using integration clustering and geostatistic method in central field of Persian Gulf. *Journal of Petroleum Science and Engineering*, *135*(1), 152–160. https://doi.org/10.1016/j.petrol.2015.08.019

Kotsiantis, S. B. (2007, July 16). Supervised Machine Learning: A Review of Classification Techniques. *October 2007*, *31*(No 3), 249–268. Retrieved from http://www.informatica.si/index.php/informatica/article/viewFile/148/140

Kumar, B., & Kishore, M. (2006). Electrofacies classification—a critical approach. (pp. 822–825). Presented at the 6th international conference and exposition on petroleum geophysics., Kolkata, India. Retrieved from https://www.spgindia.org/conference/6thconf_kolkata06/360.pdf

Kuroda, M. C., Vidal, A. C., Leite, E. P., & Drummond, R. D. (2012). ELECTROFACIES CHARACTERIZATION USING SELF-ORGANIZING MAPS. *Brazilian Journal of Geophysics*, *30*(3), 287–299. http://dx.doi.org/10.22564/rbgf.v30i3.186

Lee,S. H., Kharghoria, A., & Gupta, A.D. (2002). Electrofacies Characterization and Permeability Predictions in Complex Reservoirs. *Society of Petroleum Engineers*, *5*(03), 237–248. https://doi.org/10.2118/78662-PA

Moreira, J. L. P., Madeira, C. V., Gil, J. A., & Machado, M. A. P. (2007). *Bacia de Santos* (Vol. 15). Rio de Janeiro - Brasil: Boletim de Geociencias da PETROBRAS. Retrieved from https://d1wqtxts1xzle7.cloudfront.net/56751053/BGP_2007_15_2_60_Bacia_de_Santos.pdf?1528400761=&response-content-disposition=inline%3B+filename%3DBacia_de_Santos.pdf&Expires=1595111920&Signature=CtEBH~EsNbB34jKV7fQUtZsrnXpgzlOV31HRhGXKQ1Rv0-ApVIRcAhvQpSwOiLXik--r0uDVQjiQLWQZoZ5QTgvgUEBhuj3yMrPmONvNR8IxJMaqaHoN2zWEjNeWrGp9DsC8rkuRfHpzPgoX41SqSy6bT7BOC9NRcCkp4Z2TaNc5M5oKhj6CdxWLobOWcQXhKRqG~D9vTUrnNxUtCo8PTc~AgtdP9Y97bW5nG7zgFv2WsKuP0nYfXI4ixw8WWaOT439XlfYip2c5GNJDgcuRQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Oliveira Júnior, J. M. D. (2014). *CLASSIFICAÇÃO DE LITOFÁCIES ATRAVÉS DA ANÁLISE AUTOMÁTICA DE PERFIS ELÉTRICOS DE POÇOS DE PETRÓLEO DA AMAZÔNIA* [Master Thesis, Universidade Federal do Amazonas]. https://tede.ufam.edu.br/handle/tede/4144

Sauer, I. L. (2016). *O pré-sal e a geopolítica e hegemonia do petróleo face às mudanças climáticas e à transição energética*. São Paulo - Brasil: seesp.org.br. Retrieved from http://www.seesp.org.br/site/images/Recursos_Minerais_Ildo_Sauer_1.pdf

Sfidari, E., Amini, A., Kadkhodaie, A., & Ahmadi, B. (2012). Electrofacies clustering and a hybrid intelligent based method for porosity and permeability prediction in the South Pars Gas Field, Persian Gulf. *GEOPERSIA*, *2*(2), 11–23. https://doi.org/10.22059/JGEOPE.2012.29229

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. (pp. 1310–1315). Presented at the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India: IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/7724478

Suykens, J. A., & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *June 1999*, (9), 293–300. https://doi.org/10.1023/A:1018628609742

Vapnik, V. (1998). The Support Vector Method of Function Estimation. In *Nonlinear Modeling* (pp. 55–85). Boston, MA: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4615-5703-6_3

# Appendix C

# Neural Networks and Support Vector Machine in classification and prediction of carbonates electrofacies

# Neural Networks and Support Vector Machine in classification and prediction of carbonates electrofacies

**Andre Torres[1] and Alessandro Batezelli[2]**

[1]School of Mechanical Engineering, University of Campinas, Brazil
[2]School of Mechanical Engineering, University of Campinas, Brazil

a090388@dac.unicamp.br, batezeli@unicamp.br

**Abstract -**

**Classifying and predicting data sets has been one of the greatest challenges in current times, mainly because the amount of data acquired has increased a lot in the last years. In petroleum geology, analyzing a large volume of well log data to extract reservoir properties by manual approaches it is hard task and time-consuming. Therefore, methods and algorithms that offer to classify and predict these data have become great allies to better understand the huge amount of data and information. The Support Vector Machine and Neural Networks are algorithms that through robust mathematical calculations are responsible for auxiliary in the classification and prediction of data sets. In the petroleum reservoir characterization, the electrofacies are a theme quite discussed, mainly because they reflect the properties of the rock, fluids, and pores. Thus, geological models can be built and, consequently, they can auxiliary to minimize the uncertainties in the exploration. The work used Support Vector Machine and Neural Networks to classify and predict different electrofacies in a data set which were provided eight Geophysical well logs as input features. According to statistical validation of models, the Neural Network have presented a better performance at all scenarios when compared with the SVM Model. In the Confusion matrix evaluation method, the Neural Network Model achieved values of precision and F1-score superior than the SVM Model for all electrofacies. Besides that, the final accuracy reached by the Neural Network Model (97%) was 6% superior to the final accuracy reached by the SVM Model (91%). In the K-fold Cross Validating method, the Neural Network Model achieved a final accuracy of, 96.9% with a standard deviation of approximately 1% while the SVM Model achieved a final acurracy of, 86.6% with a standard deviation of approximately 10%. The data set of the well log predicted had two electrofacies predicted in both models, showing convergence between the models. The Support Vector Machine and Neural Network algorithms through the results obtained in the research proved to be a robust method for classifying and predicting electrofacies and a promising method capable of automating the processing of electrofacies classification. The research shows new ways to classify and predict electrofacies and tries to open up new possibilities for the use of Artificial Intelligence algorithms in the world of geosciences.**

**Keywords -**

**AI algorithms; Electrofacies; Carbonates; Brazilian Presalt**

## 1 Introduction

In the last years, Artificial Intelligence (AI) algorithms have been inserted in several areas of knowledge and the oil and gas (O&G) sector has not been left behind. In recent years the O&G has made AI algorithms more present in the industry and literature. The introduction of them in O&G have become necessary because the amount of data collected nowadays has increased at high rates and, for processing these huge amounts of data are necessary robust mathematical algorithms. Analyzing a large volume of data is required in order to develop a comprehensive understanding of reservoir distributions and their production performance characteristics [1]. The current research used two AI algorithms to classify and predict the electrofacies, they are, Support Vector Machine (SVM) [2] and Neural Networks ([3] [4] [5]). One of the most common applications of SVM and Neural Networks algorithms is to recognize and predict patterns in different data set and in this way, the SVM [6] and Neural Networks have been widely used to solve complex real-world problems [7]. Much progress has been made to understand and improve learning algorithms, but the challenge of artificial intelligence (AI) remains [8].

In geology of petroleum, an indirect way to identify the rock properties in the subsurface is through the well-logs, which are responsible for indirectly measuring of rock properties. These Geophysical well logs are able to identify the porosity, density, resistivity, and radioactivity in the reservoir. From these indirect measures of the rocks in subsurface, the Geologists are capable to recognize patterns in different packs of the rocks. For the patterns found by the expert's Geologists from the well logs are given the name of electrofacies [9]. The term represents sections of rocks in the subsurface with similar geophysical properties. Traditionally electrofacies have been identi-

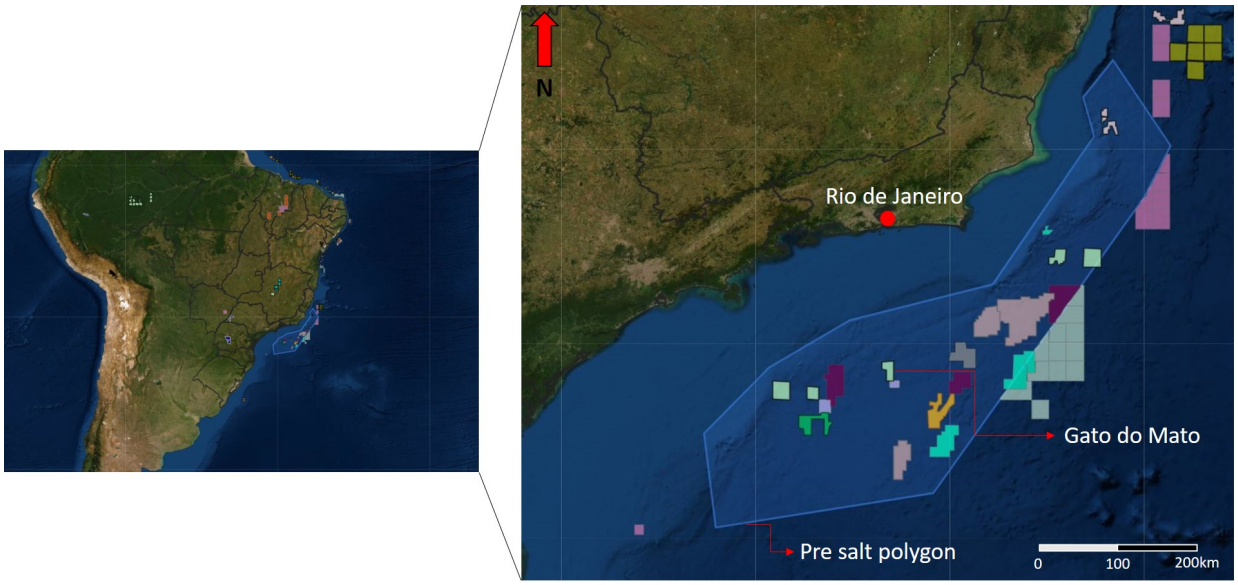



Figure 1. Localization of the Gato do Mato field in the brazilian Pre-Salt polygon.

fied manually with the aid of graphical techniques like crossplotting from wire-line logs [10]. Electrofacies have been used widely in petroleum prospecting and reservoir characterization as a tried to distinguish different beds in a petroleum field as well as in the correlation with the lithofacies [11], [12]. Most recently several AI algorithms have been introduced to try automating the task of electrofacies identification [13].

The current research was developed from carbonates reservoirs of the brazilian Pre-salt, located in Santos Basin, Gato do Mato oil field. The Santos Basin is localized in the southeast of Brazil, with an approximate area of 350.000 $km^2$ and the sediment thickness in some areas is higher than 10 km [14]. The figure 1 presents the Pre-salt distribution, as well as the study area. The brazilian Pre-salt is a province characterized by carbonate reservoirs, microbial and coquina rocks, buried at a depth that surpasses 5.000 meters, distributed in the Santos and Campos sedimentary basins, located at the southeast brazilian coast [15]. Estimates in Santos Basin suggest that the potential volume of oil reserves is higher than 100 billion barrels [16], which would position Brazil as having the fifth-biggest world reserves. The stratigraphic section studied in the Santos Basin is localized under the evaporitic unit formed during the pos rift phase in the Aptian Stage. The main stratigraphic units that compose the study area are Itapema and Piçarras Formations. The Itapema is located immediately below the evaporitic section and was formed between the Late Barremian to Early Aptian. Based on the paleogeographic distribution, in the distal portions, the Itapema Formation were formed by marine incursions that were responsible to deposited dark shales and carbonate rocks [17], while the proximal portions are constituted by conglomerates and sandstones deposited by alluvial fan [18]. The Piçarras Formation corresponds to those alluvial fan sediments, composed of conglomerates and lithic sandstones deposited during the Barremian Stage. Volcanic rocks of the Camboriú Formation (Upper Necomian) constitute the basement of the basin [18]. The carbonates reservoirs of the brazilian Pre-salt are derived from lacustrine environments, and appear to be laterally continuous over tens of kilometers [19]. The laterally continuity of them also suggest the lateral continuity of electrofacies and, in this case, the interpolation of electrofacies among the well logs presents in the Oil Field can be done with a high level of safety.

In order to identify different electrofacies related to carbonate reservoir rocks, the work proposes to use the SVM (Model I) and Neural Networks (Model II) algorithms to classify and predict the electrofacies of the Gato do Mato Oil Field, Santos Basin. The main reason to the choice by the current study area is due to the little knowledge available about the area and the carbonate rocks presents there. Last but not least the choice by the study area has the goal to become the Brazilian Pre-salt more understandable and accessible.

Due to the heterogeneity associate with the carbonates reservoirs, it is often unpredictable and hard task for associating with a pattern (carbonate electrofacies), making it a challenge for Geologists to individualize electrofacies. Thus, to predict and classify patterns in carbonate reservoirs with a high accuracy are necessary more information than in the case of siliciclastics reservoirs. To overcome this high level of heterogeneity in carbonates rocks, the

research applied a greater number of input features (8) to help the models distinguish the different electrofacies in the well log section. The current input features, Gamma-Ray, P-velocity, V-velocity, Resistivity, Water Saturation, Neutron, Sphi, and Density, were chosen according to its capacity of representation of the rock matrix, fluids, and porosity. This choice was made with the goal of understanding all elements presents in the rock interval of interest. Between the groups of Geophysical well logs able to represent the rock matrix, porosity and fluids, were chosen the Geophysical well logs in common between the well logs worked in the research. To realize the predictions were utilized the softwares, Jupyter notebook, an open-source software with Python version 3.8.3, and Google Colab with TensorFlow 2.0 version 2.5.0. The software used to extract the input features and to plot the Geophysiscal well logs section was the Petrel, version 2017. The methods utilized to evaluate the accuracy were, the Confusion Matrix and the K-fold Cross-Validating. These methods beyond that to evaluate the hit rate, they are also capable of auxiliary in the underfitting and overfitting analyses. Analyzing the methods Confusion Matrix and K-fold Cross-Validating about the accuracy is possible to note a better performance of Neural Networks (Model II) than the SVM (Model I). In the matrix confusion evaluation, Neural Networks presents a performance 6% superior to the SVM Model (Model I), and in K-fold Cross-Validating the difference achieves 10%.

The current research has been contributed to the discussion about the importance and value of AI algorithms in electrofacies classification and prediction. In addition, the research has shown the challenges faced when working with carbonate rocks. Once the challenges involved in this process, have been overcome, the results achieved can guarantee more precision and agility in electrofacies classification. Therefore, due to the high impact to literature and Petroleum Industry it is crucial to make more visible the theme AI algorithms in O&G. The next step to the future is taken every day with the advance of knowledge in different areas. At the moment, it is unquestionable that AI algorithms are the bridge to the faster development of many areas of knowledge. Thus, it is clear that AI algorithms can be a great partner in this long journey to the future.

## 2 Development

The research worked with the SVM and Neural Networks algorithms with the goal of classifying and predicting electrofacies of a data set created from Geophysical well logs. The raw data set (primary data) of the research was provided by the National Agency of Petroleum from Brazil. The features chosen by research to compose the data set were Gamma-Ray, P-velocity, and V-velocity to inform about the rock properties. Resistivity and Water Saturation to know the information about the fluids. Neutron, Sphi, and Density to obtain information about the porosity. These eight Geophysical well logs were used as input features in the models applied in the research. They were chosen with the goal of representing the triple combo, rock, fluids, and porosity. Among the geophysical well logs provided to the research, these were the features presents in all well logs simultaneously. The other Geophysical well logs able to represent the triple combo(rock, fluids, and porosity) were not present simultaneously in all well logs. How will be explained forward, in both algorithms, SVM and Neural Networks, are necessary a labeled data set in which the algorithm will learn how to make the classification, in other words, the external user needs to inform the algorithm what will be classified and predicted. One way of teaching algorithms the classification and prediction is labeling the data set, in the research this step was made by experts Geologists and inserted to the models through the data set labeled. Labeling the data set in the research case is equivalent to making a manual electrofacies classification. Thus, according to the Geophysical well logs chosen to compose the data set, the electrofacies applied in the research represent the measurement of the three properties of the reservoir and non-reservoir zones, rock, fluids, and porosity. Once electrofacies are defined as a merge of these three properties, the electrofacies will represent reservoir and non-reservoir zones rather than represent lithofacies. The main difference between electrofacies and lithofacies is that the first represent a Geophysical well logs response of the rock, pores and fluids, while the second is defined as a rock unit composed of distinct lithological features, including composition, granulometry, and sedimentary structures [11], [12]. Therefore, it is important to understand that electrofacies and lithofacies will not always match each other. In order to know more about the physics proprieties of the rocks is necessary the auxiliary of the petrophysics, it is a more precise way to determine the physical properties of the rock from microscopic analyses of the rock [20].

The current research was developed from carbonates reservoirs, thus, it is important to remember the big difference in classification between carbonates and non-carbonates rocks. Carbonate rocks have a set of distinct characteristics that make them unique in geological studies. Several classifications for carbonate rocks have been proposed [21], [22], [23]. In some classifications the main features used to classify are the presence or absence of matrix versus cement [21], [22]. Other classifications focused in texture, amount of lime mud and abundance of grains [23], this classification is also able to reflect the energy level in the environment. Due to this big difference between carbonate and non-carbonate rocks, the research

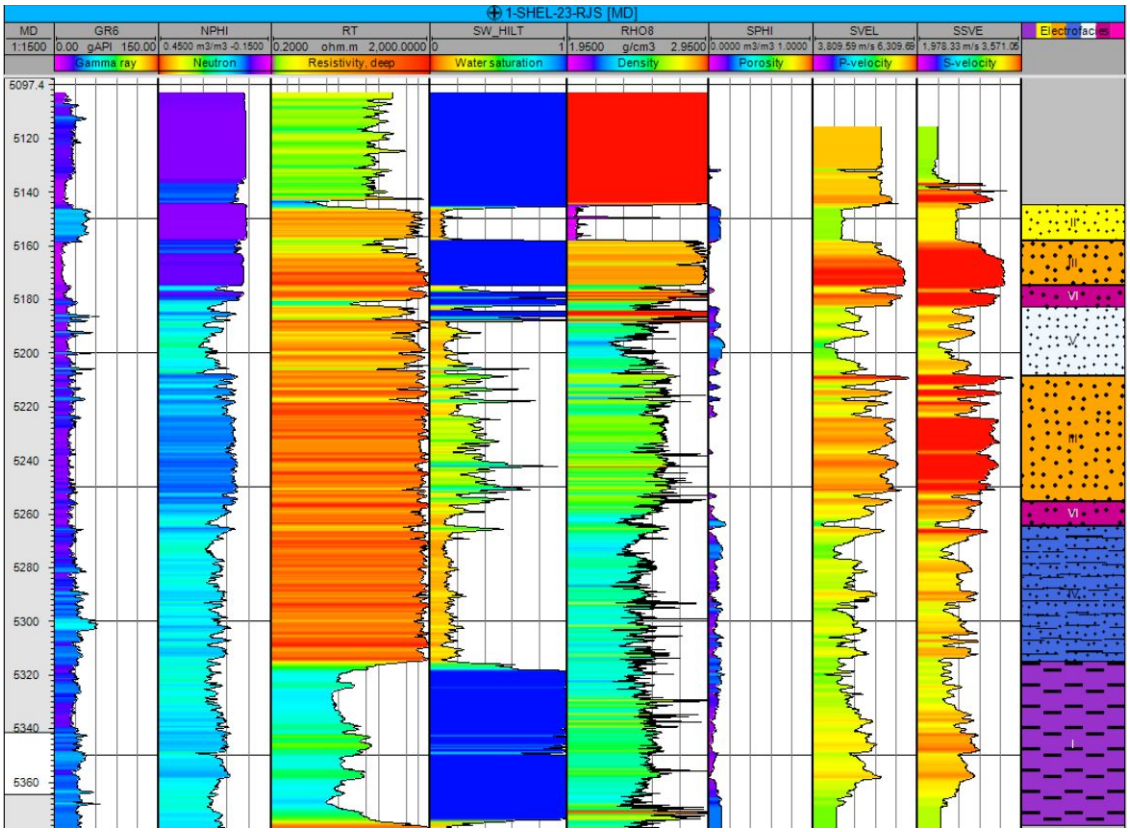



Figure 2. Well log section of the well log reference, 1-SHEL-23-RJS.

has a great challenge in working with the prediction and classification of carbonate electrofacies. Mainly because the carbonate rocks have a high variation in their composition and texture, which in many cases becomes a hard task to note in the Geophysical well logs. In order to solve this problem, the research proposes the use of a data set with properties of the rock, fluids, and porosity allied with AI algorithms to distinguish the high variation in the composition of carbonate rocks. The first stage in the development of the current research involves the selection of samples corresponding to the intervals of interest. In the current research, the interval of interest corresponds to the Brazilian pre-salt, a region rich in oil and gas where are located the carbonates reservoir.

In the study area (Figure 1), Gato do Mato oil field, the carbonates reservoir are located at depths greater than 5.000 meters. The interval of interest has up to 400 meter and presents geophysical and lithological characteristics of the carbonate rocks. After selecting the interval of interest, data pre-processing,and data science stage, was plotted the Geophysical well logs (eight input features) in a well log section. This procedure was made with the help of the Petrel software, version 2017. With the well log section was possible to make a manual classification of the 1-SHEL-23-RJS data set. The manual classification (Figure 2) was made in order to become the data set labeled. Supervised learning algorithms are only able to learn after the data set is labeled. With the data set labeled was necessary to realize the validation of the model. The validation of the model was made splitting the data set in two data set, training data set and test. Analyzing the data set, it is possible to verify the accuracy of the model and whether the model is in the case of overfitting or underfitting. The model was considered able to make prediction when was verified non overfitting or underfitting, and an accuracy above 85%. The workflow described above can be seen in the flowchart illustred in the figure 3.

## 3 Conventional Workflow

The conventional process to classify electrofacies is made using Geophysical well logs that represent different rock properties in the subsurface. From the response of the Geophysical well logs, the specialized Geologist beginning the identification process of the electrofacies. To make the electrofacies classification the specialized Geologist verify the patterns present in the Geophysical well logs and then individualize each of them. This process is repeated to all wells present in the petroleum field. If the

petroleum field is huge and there are many well logs in the field, then, the manual electrofacies classification can become a time-consuming process.

The manual classification shows to have unless two disadvantages, (i) a time-consuming process, in the case that petroleum field has a lot of well logs to be analyzed, (ii) the classification can change according to the specialized Geologist. In the second (ii) case the trouble is more serious because in this case, the electrofacies classification will not have a standard and consequently leads to one subjective interpretation of electrofacies.

## 4 Proposed Workflow

The proposed workflow was thought with the goal of becoming the electrofacies classification an automated process, and then, save time in the process of electrofacies classification. To achieve the main goal, the research proposed the application of Artificial Intelligence (AI) algorithms to make the classification and prediction of electrofacies in Geophysical well logs of carbonate rocks. This process usually succeeds in achieves a high hit rate and ensures a standardized in the electrofacies classification and prediction. Thus, the research proposed initially a manual classification in Geophysical well logs of only one well (1-SHEL-23-RJS). This well was called in the research of reference well (Figure 2). The manual electrofacies classification is a necessary process because it allows to labeled the data set of the Geophysical well logs. Both algorithms utilized in the current work, Support Vector Machine and Neural Networks require a data set labeled to train, learn, and posteriorly making the predictions. From the data set labeled, the models that were built to predict electrofacies in a new data set (unknowns data set) are able to work normally. This happens because, when the reference well is manually labeled, the specialized Geologist is teaching the algorithms applied in the models how the electrofacies classification must be done accurately to prevent big errors and lack of standards. The proposed workflow can be seen in the figure 3.

## 5 Support Vector Machine

The Support Vector Machine is an algorithm that was developed by Vladimir Vapnik in 1960s, and in the nineties the algorithm was highlighted to solving pattern recognition problems through of the Vapnik's works ( [24]; [25]). The SVMs method maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space. The algorithm works with a concept of maximizing the minimum distance from hyperplane to the nearest sample point [26]. This basically involves solving a quadratic programming problem [27]. This technique was first developed for the restricted case of separating training data without errors, later was enhanced the case of separating data with erros. Thus, there are two historic Support Vector Machine algorithms, (i) the hard margin Support Vector Machine and (ii) the soft margin Support Vector Machine that will allow for an analytic treatment of learning with errors on the training set [28]. In both algorithms the elements, Hyperplane, Margin and Support Vectors are essential. The hyperplane is understood as a plane that separate the data, the margin is the distance of the hyperplane, and the support vectors are the closest points of the hyperplane. The optimal hyperplane is defined as the linear decision function with maximal margin between the vectors of the two classes [28]. According to [28] to find the optimal hyperplane in the hard margin case the algorithm considers the set of labeled training patterns:

$$(y_1, x_1), ..., (y_l, x_l) \quad y_i \in -1, 1 \tag{1}$$

The data set is linearly separable if there is a vector W and a scalar b such that

$$y_i(W \cdot x_i + b) \geq 1 \quad i = 1, ..., L \tag{2}$$

are valid for all elements of the training set. The optimal hyperplane is given by

$$W_0 \cdot x + b \tag{3}$$

The maximal distance between the projections of the training vectors of two different classes under the constraints (2) will ensure an optimal hyperplane (3)

$$\varphi = W \cdot W \tag{4}$$

for this the equation (4) has to be minimized. To do this a standard optimization technique is used and a Lagrangian is constructed

$$L(W, b, \Lambda) = \frac{1}{2}(W \cdot W) - \sum_{i=1}^{L} \alpha_i [y_i(x_i \cdot W + b) - 1] \tag{5}$$

where $\Lambda^T = (\alpha_1, ..., \alpha_l)$ is the vector of non-negative Lagrange multipliers corresponding to the constraints (2). To find the hyperplane in the soft margin case the reasoning is similar and can be found in the article of Cortes and Vapnik (1995)[28].

The SVM algorithm deal very well with linear data sets, in cases where a straight line is sufficient to separate a data set, but for problems with non linear data sets are not possible to find a straight line to classify the data set, in this case the Kernel Trick [29] is used. The basic idea is that if a data set is inseparable in the current dimensions, the kernel trick [30] will carry the input data set to a higher dimensional space and by choosing an adequate

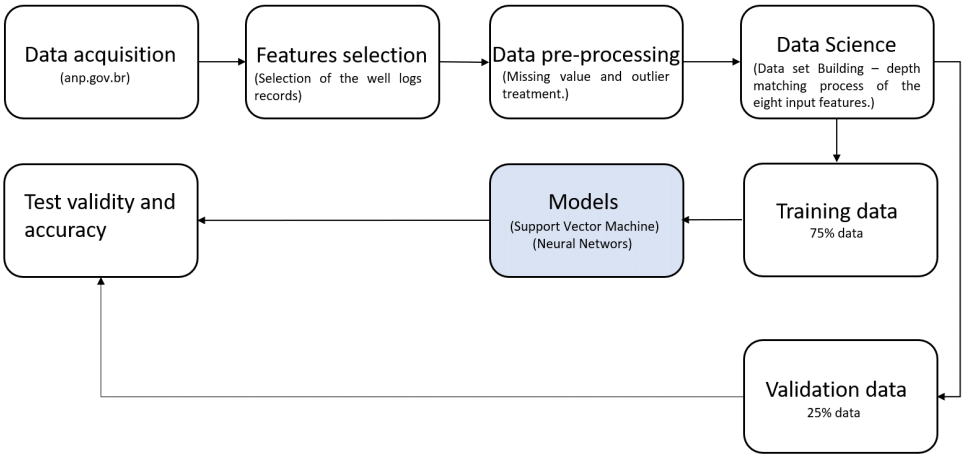



Figure 3. The flowchart of the proposed workflow.

dimension, the data set points become linearly separable or mostly linearly separable in the high-dimensional space [31]. Thus, the SVMs with the Kernel trick currently has been ensured the solving of linear and non linear problems efficiently.

## 6 Neural Networks

Neural Networks are a field of Artificial intelligence (AI) in which mathematician algorithms have been worked in a process that remember the brain thinking process. Neural Network models are biologically plausible and can help to understand how the brain works [32]. The elements that compounds a Neural Network are illustrated in the figure 4, its shows a biological neuron on the left and an artificial neuron on the right. There are some points to pay attention:

(i) The first point are the input nodes, is through them that the Model is feed with the input data, in recognize problems this nodes can be understood as the object features;

(ii) Each connection between the input nodes and the hidden layer nodes has a weight associated with connection, based on the individual weight of each connection the algorithm selects which input features will have greater or lesser impact on processing;

(iii) The weighted sum is a calculus that can be represented by the equation 6, to calculated the weighted sum is necessary, the connection weight and the input node value. The weighted sum is expressed by the equation below:

$$\textbf{Weighted Sum} = b + \sum_{i=1}^{n} X_i * W_i \qquad (6)$$

Where $X_i$ is the input node value, $W_i$ the weight associated to connection and $b$, bias, an additional weight that allows to move the activation function to the left or right to improve Model learning;

(iv) The result obtained with the weighted sum is transfer to activation function. The activation function is chosen according to model goal;

(v) The output node is the last compound of the default Artificial Neuron, it is function of the all elements listed before and can be expressed as follow in equation 7:

$$\textbf{Output node} = Y = f(b + \sum_{i=1}^{n} X_i * W_i) \qquad (7)$$

The process described above can be understood as feed-forward, which is the information flow in the Neural Network from input data to the output data [33]. The feed-forward doesn't allow the update of the weights, because this is necessary the application of the back-propagation. The discovery of "back-propagation" in the context of Neural Networks by Rumelhart [34] drastically improves the learning efficiency of such models enabling them to be used in practice [35].The central idea of error back-propagation is to compute the partial derivatives of the weights in a Neural Network by applying the chain-rule repeatedly [35].

The learning rate can be understood as the amount that is updated in each epoch in the weights of the Model. It is used to find the best combination of weights so that the Model reaches the minimum error in its predictions. The procedure of choosing a good value for the learning rate of the stochastic optimization can increase performance and reduce the training time in the Model. But if a value is chosen for the learning rate bigger than expected, the Model can update the weight more than necessary and occasionally jump the minimum value of error, in this case, the Model will have a needless time consuming. On the

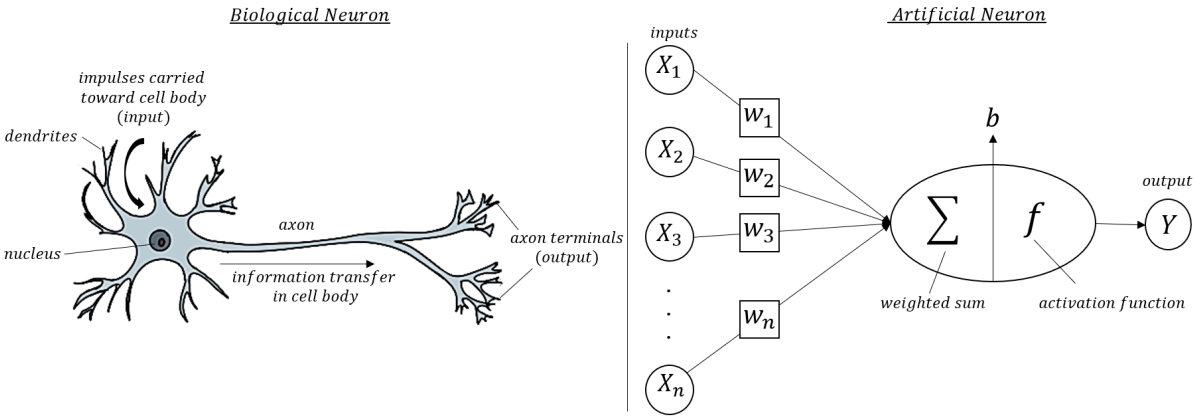



Figure 4. Biological and artificial neurons showing the similar compounds.

other hand, however, if the value for the learning rate is less than expected, the Model will have to update several times until it reaches the minimum error value, and again this process will be unnecessarily time-consuming. The learning rate can be considered the most important hyperparameter, thus, choosing a good value for the learning rate it is fundamental in training deep learning Neural Networks.

The main purpose of the activation function is to adds the nonlinear factors to remove redundant data while preserving features, it retains "active neuron feature" and maps out these features by nonlinear functions, which is the essential of the Neural Network to solve complex nonlinear problems [36]. The Relu (Rectified Linear Units) activation function ([37], [38], [39]) is a mathematical function whose output 0 if the input is negative, and for any positive value x, it returns that value back. In Neural Networks modeling of electrofacies, predicting the probability of the electrofacies requires computing scores for every electrofacies in the data set and to normalize them to form a probability distribution. This is typically achieved by applying a softmax function. The Softmax activation function ([40], [41]), is often used in the output layer of Neural Network Models for multi-class classification problems, where the number of output classes is required on more than two class labels. Softmax makes this possible because the output is a vector with the probabilities of occurrence of each class label.

The hyperparameter, batch size, used in the Model corresponds to the number of training samples that will be executed before the weights update. The advantages of using a batch size smaller than the number of samples are:

(A) Less memory is needed. When the data set is very large this is a crucial factor;

(B) Normally, Neural Network train is faster, in addition, Neural Network parameters will have more than just an update because the Model will have updated Neural Network parameters after each batch size. In the case that, the batch size is equal to the sample number, the Model will only have an update of the Neural Network parameters.

Several recently proposed stochastic optimization methods have been successfully used in training deep networks such as RMSPROP, ADAM, ADADELTA, NADAM. They are based on using gradient updates scaled by square roots of exponential moving averages of squared past gradients [42]. In the Model applied in the research, the optimizer chosen was the Adam optimizer due to its lower training cost and faster convergence in relation to the other optimizers. The name Adam derived from adaptive moment estimation. Adam is a method for efficient stochastic optimization that only requires first-order gradients with little memory. The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients [43]. Therefore, Adam is an optimizer that can be used as a good alternative to the classic Stochastic gradient descent, the Adam optimizer is called when upon updating Neural Network weights.

In spite of many successes, Neural Networks still suffer from a major weakness. The presence of nonlinear hidden layers makes deep networks very expressive models which are therefore prone to severe overfitting [44]. Due to this reason, was applied the dropout technique to the Neural Networks Model used in the research. The dropout is a technique used in Neural Networks to try to avoid the overfitting of the algorithm. The key idea is to randomly drop units (along with their connections) from the Neural Network during training [45]. Besides to avoid the overfitting, the technique called dropout has shown

to significantly improve the performance of deep Neural Networks on various tasks [46].

## 7 Materials and Methods

The research used data from well logs present in a Brazilian Pre-salt oilfield, Gato do Mato. The data set of the Geophysical well logs of each well was provided by the Brazilian National Agency of Petroleum (ANP). The algorithms applied to the models were developed in Python programming language. The research has built two models with accuracy level greater than 90%, Models I and II, The environments Google Colab, Jupyter Notebook, and the software Petrel version 2017 were used to develop the research. The Model I (SVM) was built with the help of the Scikit-learn Python library, version 0.24.2 in the Jupyter Notebook environment. The Model II (Neural Networks) was built with the help of the TensorFlow 2.0 version 2.5.0 in the Google Colab environment. The link between the electrofacies predicted by the models and the well logs section was made with the Petrel software, version 2017.

### 7.1 Data set

The building of the data set was made using the data provided by ANP. For the building of the data set the first step was selecting the features that were used in the models. As the main objective of the research is the classification and prediction of electrofacies, the Geophysical well logs used were based on three proprieties of the Formation, rock, porosity, and fluids. There are several Geophysical well logs capable of measuring these three properties (rock, porosity, and fluids), but to avoid problems of Geophysical well logs absence in the data set were used Geophysical well logs in common between the wells worked in the current research. The Geophysical well logs used were, about the rock: Gamma-ray; P-velocity; and V-velocity. About the porosity: Neutron; Sphi; and Density. About fluids: Resistivity and Water Saturation. These are the eight input features responsible for informing about the properties of the Formation. Initially, the P-velocity input feature was chosen, but due to the linear dependence (Figure 9) between P-velocity, and S-velocity, the research opted by removed the P-velocity in order to optimize the algorithm.

### 7.2 Data science stage

As a core of the research, the data science stage was concentrated in the creation and treatment of the data set. In the current research the term data science was used to describe the process of creation and treatment of the data set. The building of the data set is an essential stage and must be done with expertise because it is the guarantee of reliable information. In the data science stage was realized the depth matching process, and statistical processing in the data set.

It is important to remember that, the initial efforts to work with the data set at the data science stage, are the core of researches related to the application of AI algorithms. When the building, and treatment of the data set is well-realized, AI algorithms application become more efficient, in some cases can increase the hit rate and consequently lead to better results. The evolution of the research in models I and II just was possible to achieve due to the precise development realized in the data science stage.

#### 7.2.1 Depth matching process

The first challenge in the data science stage is related to the match of the eight input features in-depth. Each input feature is provided individually and many of these Geophysical well logs do not have records at the same depth, thus creating an algorithm capable of building a data set with the eight matching records in depth is a crucial step for the well succeeding progress of the project.

The depth-matching process of the eight input features in relation to depth was responsible for a significant reduction in the number of input records. Initially, the number of Gamma ray measuring collected was 7277, and the number of Density records was 11707. After the application of the depth-matching algorithm in the eight input features, the number of records decreased to 1624 input records.

#### 7.2.2 Statistical processing

After the depth matching process, the data set was normalized in order to transform all geophysical well logs in the interval between 0 and 1. This is an important process because it prevents that the absolute values of the records have a weight higher or smaller. In order words, the normalization process makes the model training less sensitive to the scale of input features. Initially were chosen eight input features, these input features were normalized and their frequency histograms were plotted as can be seen in the figure 5.

The procedure for the building of frequency histograms was made due to the outlier detection method, when the data distribution has a normal distribution behavior it is common to use the method mean plus or minus two/three standard deviations [47]. In this method the values that are outside of this range are considered outliers. When it is used the mean plus or minus two standard deviations, 95.45% of the data are present in the range. The method of the mean plus or minus three standard deviations is based on the characteristics of a normal distribution for which 99.87% of the data appear within this range [48]. But, it is important to remember that these results just are valid in the normal distribution case. Beyond of

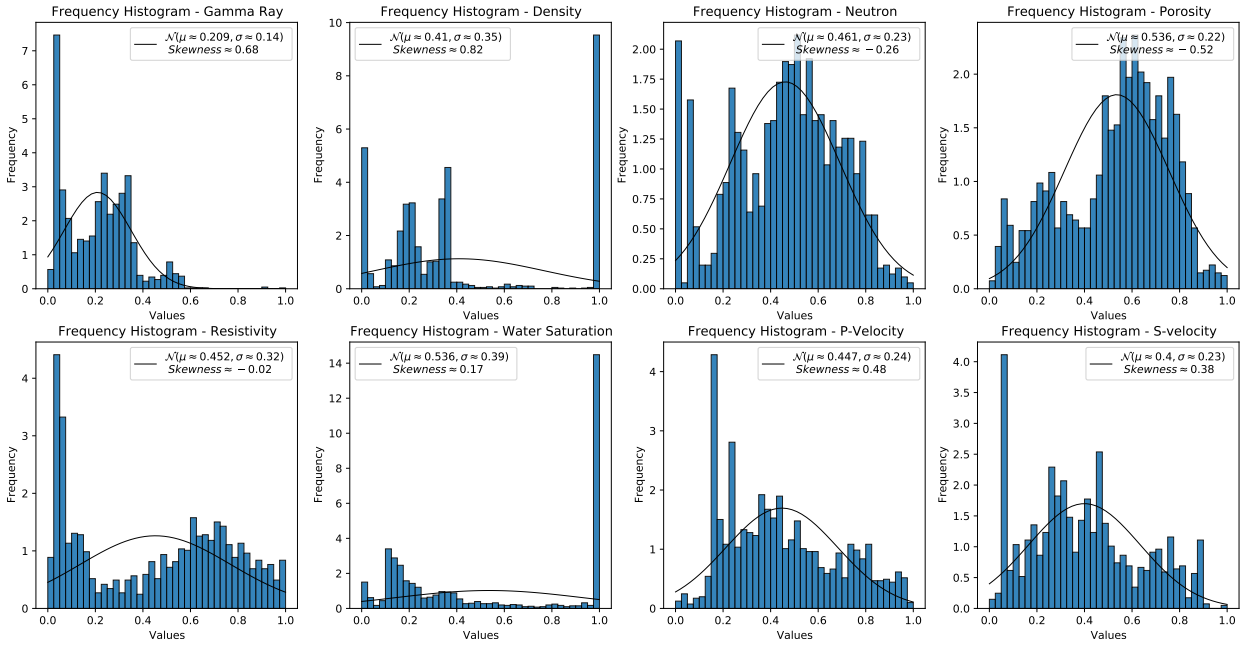



Figure 5. Frequency histogram of the Geophysical well logs. Data set normalized.

the visual aspect of the frequency histograms, the research also calculated the Fisher-Pearson coefficient of skewness to auxiliary in the analysis. In the figure 5, frequency histograms are plotted with the calculus of the mean($\mu$), standard deviation($\sigma$), and the Fisher-Pearson coefficient of skewness.

The sample skewness is computed as the Fisher-Pearson coefficient of skewness [49]. The Fisher-Pearson coefficient of skewness is responsible to measure the lack of symmetry in a distribution. Normal distribution has skewness 0. Larger values (in magnitude) indicate more skewness in the distribution of observations [49]. Thus, the Fisher-Pearson coefficient of skewness it is useful to help in detection of normal distribution. The Fisher-Pearson coefficient of skewness was applied in the eight input features, and the values can be seen in the figure 5. It is possible to note from figure 5 that the population of the input features have unknown distributions, and with a notable asymmetry (skewness). Thus, the method of outliers detection based in the mean plus or minus two/three standard deviations is not valid to use in the research data set. Due to the notable asymmetry and the unknown distributions illustrated in the figure 5 was necessary to find another method capable of detecting the outliers with reliability.

One first approach would be to make the data symmetrical by the use of a non-linear transformation, in the research case, the logarithmic transformation was used. Presumably with such an approach, outliers would also be more symmetrically away from the central tendency, providing an equal chance of locating low and high outliers. However, there is no single technique that makes the data symmetrical when they have been contaminated with outliers [50]. How can be seen in the figure 6, the logarithmic transformation does not have successful in to become the distributions normal. But, in general, the values of skewness of the frequency histograms have decreased. The exceptions were the Resistivity, Water Saturation, and P-Velocity distributions. These input features had the absolute magnitude of skewness values increased.

Once that, the logarithmic transformation was not capable to transform the distributions as normal distributions, the method Inter-Quartile Range (IQR) method of outlier detection was applied [51], [52], [53]. The method Inter-Quartile Range (IQR) method of outlier detection is usually used when the distributions are not normal and asymmetrical. The Inter-Quartile Range (IQR) is the difference between third and first quartile. Quartile are responsible to divide the ordered sample observations into four quarters having the same number of observations (m) in each quarter [54].

The skeletal boxplot consists of a box extending from the first quartile ($Q_1$) to the third quartile ($Q_3$); a mark at the median; and whiskers extending from the first quartile to the minimum $Q_1 - (1.5 \cdot IQR)$, and from the third quartile to the maximum $Q_3 + (1.5 \cdot IQR)$ [51].

The bloxplots of the distributions were created in Jupyter notebook and Google colab 7. In the method Inter-Quartile Range (IQR), method of outlier detection, the samples that are outside of the interval between the minimum $(Q_1 - (1.5 \cdot IQR))$ and the maximum $(Q_3 + (1.5 \cdot IQR))$

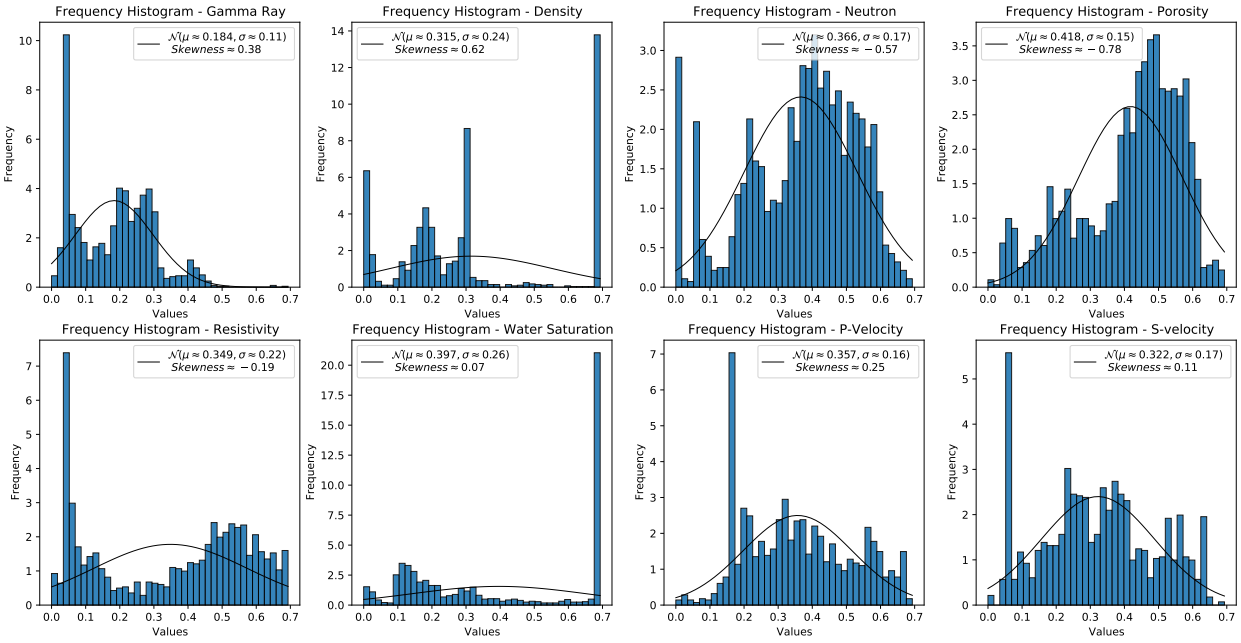



Figure 6. Frequency histogram of the Geophysical well logs. Data set normalized and log-transformed.

are considered outliers.

It is possible to note in the figure 7, that boxplots of Gamma ray and Porosity distributions has presented outliers respectively above the maximum value, and below the minimum value.

Once the outliers were detected, an algorithm created in the research was executed to remove the rows from the data set that contains the outliers. Before to remove the outliers, there were 1624 records in the data set, and after the outliers removal the numbers of records in the data set decreased to 1617 records. Among the removed data, three were porosity outliers, and four were gamma ray outliers. The porosity outliers were detect below to the minimum value in the boxplot, while the gamma ray outliers were detect above to the maximum value in the bloxplot. The final data set without the outliers can be seen in the figure 8.

After detecting the outliers, there is only one more step to be made to become the algorithm more optimized. The last step it is to verify if the input features are independent between them self. This important step avoids that the model working with redundant data (dependent variables). In this step it is calculated the Pearson correlation coefficient [55]. If the Pearson correlation coefficient is 1, the variables are totally dependent on each other. If the Pearson correlation coefficient is -1, the variables will have a total inverse dependence. In the other side, if the Pearson correlation coefficient is equal 0, the variables are independent or uncorrelated. In the research, the Pearson correlation coefficient of all variables was calculated, then with the values found a matrix was built, this matrix with all values plotted is known as correlation matrix. The correlation matrix of the variables (input features or geophysical well logs) can be seen in the figure 9.

From the figure 9 it is possible to note that the correlation matrix between the geophysical well logs presents a linear dependency between the two variables P-velocity and S-velocity. The value of the Pearson correlation coefficient for these two variables it is equal 1. Thus, for optimizing the algorithm one of these variables needs to be removed. In the research case was removed the variable P-velocity. The correlation matrix after the removal can be seen in the figure 10

Therefore, after the statistical processing stage (last stage in "Data science stage"), the data set it is finalized and ready to be used in models.

One difficulty with treatments of outliers is that there is no unanimously accepted theoretical framework for the treatment of outliers. Various fields have developed various approaches and rare are the approaches that can be formulated with the concepts of another approach [50]. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation, and incorrect results. Therefore, it is important to identify them prior to modeling and analysis [56], [57], [58]. In the research, the Inter-Quartile Range (IQR) method was chosen, which showed, through boxplot analysis , a good adaptation for detecting outliers in the data set.

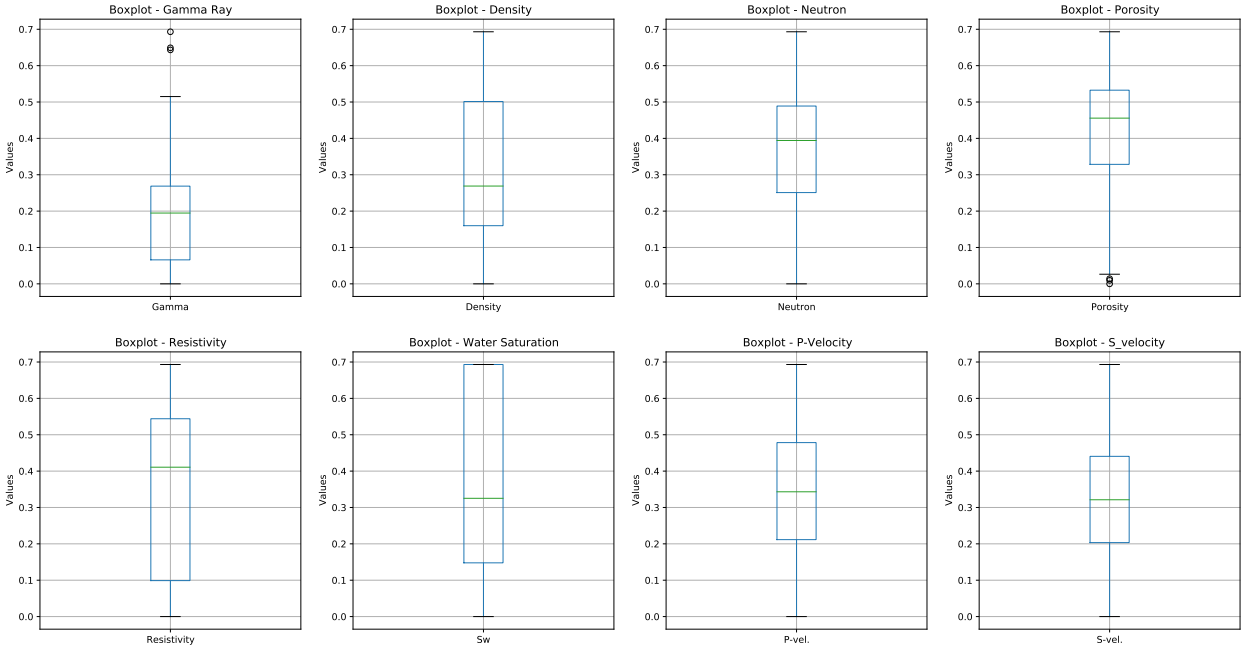



Figure 7. Boxplot of the Geophysical well logs. Data set normalized and log-transformed.

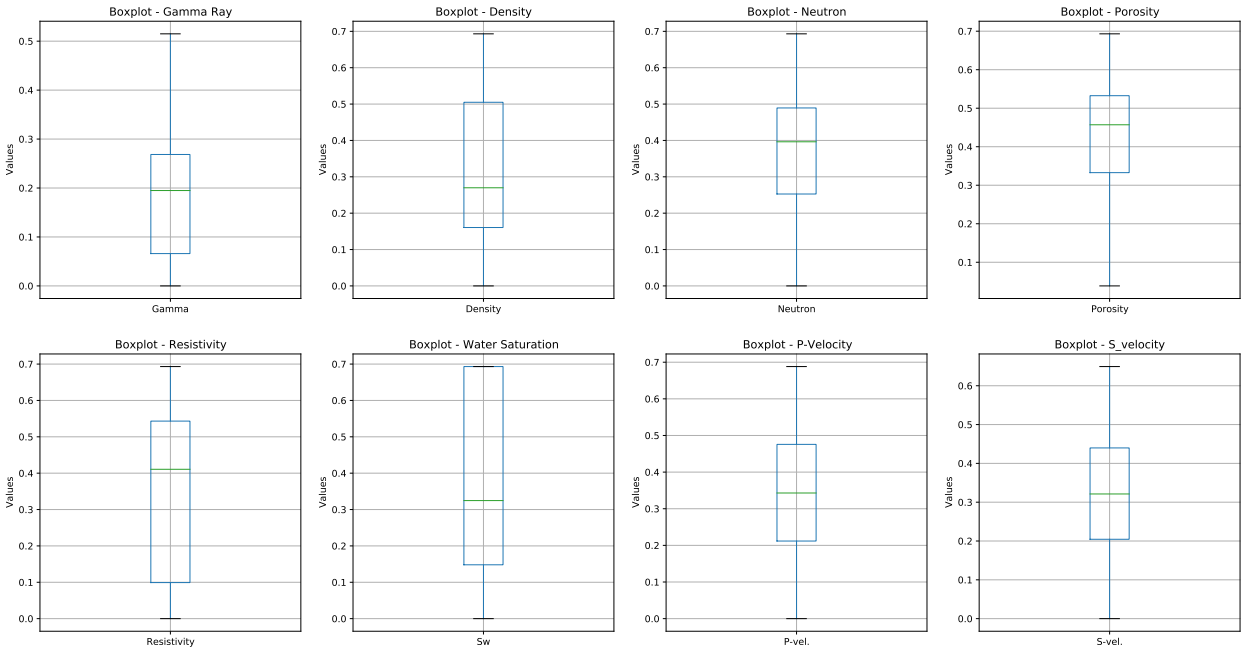



Figure 8. Boxplot of the Geophysical well logs. Data set normalized and log-transformed and outliers removed.

### 7.3 Training and Test data set

The training and test data set were divided respectively in the proportion of 75% and 25%, in the 1-SHEL-23-RJS data set, the division was made at random and proportional to each electrofacies labeled in the manual classification stage. The main data set, 1-SHEL-23-RJS, was created according to a manual classification made by a specialist Geologist. This classification was responsible for the labeling of the data set. The labeling process is necessary because it allows to evaluate the performance of the models for supervised learning algorithms. The training data set is present only in the 1-SHEL-23-RJS data set. In the Models I and II, the training data set were used to teach

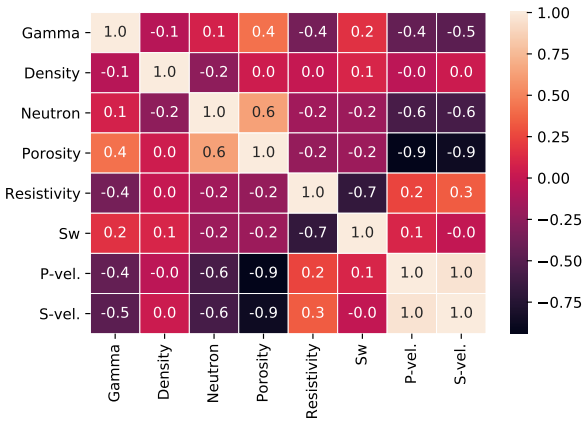



Figure 9. Matrix Correlation between the Geophysical well logs. Data set normalized and log-transformed and outliers removed.

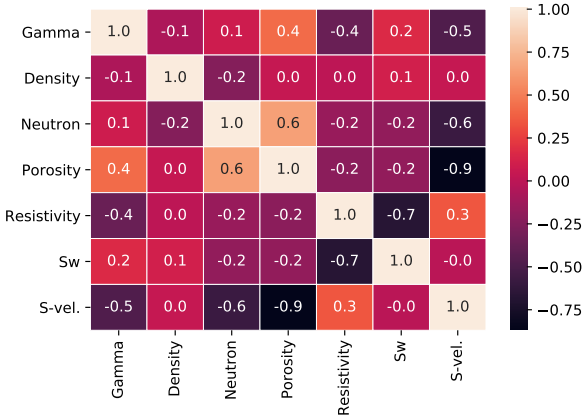



Figure 10. Matrix Correlation between the Geophysical well logs without P-velocity. Data set normalized and log-transformed and outliers removed.

the models, while the test data set present in the 1-SHEL-23-RJS were used to make the statistical validation.

### 7.4 Models

The Model I was formulated using the SVM algorithm, and was developed in Python language, version 3.8.3, with the auxiliary of the Jupyter Notebook. It was built using the Scikit-learn [59], a Python library, version 0.24.2. The data set built in the first stage, data science stage, was utilized for application of the SVM. The Model used the following hyperparameters setup, $C = 100$ (the penalty of the error), $Kernel function = linear$ (the Kernel functions available in the library are, 'linear','rbf', 'poly' and 'sigmoid'), $degree = 3$ (Ignored by linear kernel function), $gamma = scale$ (it is the Kernel coefficient for

Table 1. Number of samples of each electrofacies selected to train models I and II.

| Electrofacies | Proportion of Samples |
|---|---|
| Electrofacies 0 | (125/507) 0.246 |
| Electrofacies 1 | (025/087) 0.287 |
| Electrofacies 2 | (125/532) 0.235 |
| Electrofacies 3 | (087/333) 0.261 |
| Electrofacies 4 | (044/165) 0.267 |

'rbf', 'poly' and 'sigmoid' Kernel functions. Ignored by linear kernel function), $coef0 = 0.0$ (Independent term in kernel function. Insignificant in linear kernel function), $shrinking = true$ (it is a parameter that auxiliary the convergence velocity of the algorithm, decreasing the training time), $probability = false$ (when enabled, probability estimates are calculated), $tol = 0.001$ (it is a stopping criterion), $cache\_size = 200$ (the size of the kernel cache in MB), $class\_weight = none$ (set the parameter $C$ of class i. If not given, all classes are supposed to have weight one), $verbose = false$ (when enabled return a verbose output), $max\_iter = -1$ (number maximum of interactions, -1 for no limit), $decision\_function\_shape = ovr$ (it is the decision function present in the classifier), $break\_ties = true$ (if true, it is responsible for breaking ties according to the confidence values of the decision function), $random\_state = 3$ (it is responsible for controls the pseudo-random number generation for shuffling the data for probability estimates). Although SVM has a few tuning parameters, they have to be carefully chosen to obtain good results. To help in choosing the best hyperparameters of the Model was used the Grid Search ([60]; [61]), a method of parameter optimization used on discrete sets of hyperparameters to select the optimal ones with the aid of cross-validation [60]. This method basically test all parameters combination and measuring the efficiency according to the metric chosen by the developer. In the current research, accuracy was chosen as the metric to measure the performance of all hyperparameters combination. Thus, the hyperparameters combination that returned the highest value of accuracy was used to set up the models. The hyperparemeters that were utilized to feed the Grid Search were, $C$ (the penalty of the error), and $tol$ (it is a stopping criterion), both hyperparameters with input values of: 0.001, 0.01, 0.1, 1.0, 10, 100, 1000. The optimal hyperparameters values returned by the Grid Search were $C = 100$ and $tol = 0.001$.

In the Model II was applied a Neural Network to classify and predict the Geophysical well logs. The Neural Network applied to the Model II has eight input neurons, two hidden layer, a dropout [45] of 0.2 between the hidden

layers, and five output neurons. The first hidden layer has 128 neurons, and used the relu activation function. The second hidden layer has 32 neurons, and used the softmax activation function. The optimizer chosen was the ADAM [43] due to the best results obtained. The Model II was developed in Google Colab environment using the TensorFlow 2.0, version 2.5.0, the main library used to implement the Neural Network was the Keras, version 1.1.2. The Model II is represented by the figure 11.

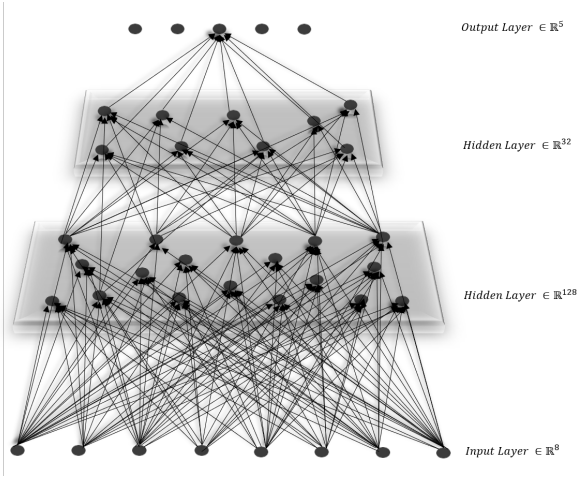



Figure 11. Representation of the Model II. The Neural Network applied in the research.

### 7.5 Statistical Validation

Estimating the accuracy of a classifier induced by supervised learning algorithms is important not only to predict its future prediction accuracy, but also for choosing the best classifier from a given set of models [62] [63]. In the work was used the Confusion matrix [64], [65], [66] and K-Fold Cross-Validating [67], [68], [69] to evaluate the Model. The Confusion matrix is built by plotting the predicted and real data in the axis x and y of the Table, in a multi-class Confusion matrix, the element $N_{i,j}$ present in the Confusion matrix is called True positive when the predicted data is equal to the real data, this situation is verified when the row i is equal to the column j ($i = j$). There are four elements in the Confusion matrix responsible for statistical validation, they are accuracy, precision, recall, and F1 score. The accuracy can be defined mathematically as:

$$\textbf{Accuracy} = \frac{\sum_{i=1}^{C}(TP)_i}{\sum_{i=1}^{C} N_i} \tag{8}$$

where $TP_i$ are the true positives, the positive data that were correctly predicted in the class $C_i$, C is the number of classes, and $N_i$ is the number of samples in class $C_i$.

The precision evaluates inside the number of samples that were predicted positive for one class, how many really are positive to the class predicted. The precision evaluation can be described as:

$$\textbf{Precision}(C_i) = \frac{TP_i}{TP_i + FP_i} \tag{9}$$

where $C_i$ indicates the class that measure was taken. $FP_i$ represents the false positive, the samples that were predicted to be positive, but in fact, are negative.

The recall is the fraction of samples predicted as true positive inside all positives in the class. The recall can be understood as the number of samples positive that were correctly predicted as positive by the Model. The recall is defined as:

$$\textbf{Recall}(C_i) = \frac{TP_i}{TP_i + FN_i} \tag{10}$$

where $FN_i$ are the false negative, samples that were predicted to be false, but in fact are positive.

The F1-score is a combination between precision and recall. Mathematically the F1-score can be understood as a harmonic mean of recall and precision and, in some cases, can be considered a measure more representative than accuracy.

$$\textbf{F1-Score}\ (C_i) = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{11}$$

The other technique used to evaluate the model performance was the k-Fold Cross-Validating, the main use of this technique is to verify the capacity of generalization of the Model and to estimate the prediction accuracy [67], [70]. This method consists of splitting the dataset into k subsets of equal or nearly equal sizes, where each subset is known as a fold and is stratified: that is, each fold attempts to retain the same class distribution. A Model is then trained with k - 1 folds and tested on the remaining fold. This process is repeated k times until all sets have been used for testing once. Thus, the experiment returns k estimates of the Model classification error [69].

## 8 Results

The research achieved some potentials results in the prediction of electrofacies of the well 1-SHEL-26-RJS. In the prediction of electrofacies, the two algorithms used in the research (SVM and Neural Networks) were responsible to predict the same electrofacies, purple and orange (Figures 12 e 13) indicating a convergence of the models I and II.

The Confusion matrix and Classification report generated in the Model I are illustrated in Tables 2 and 3:
where, $EF = electrofacies$ and $Spp = support$, the number of samples that were used in the evaluation.

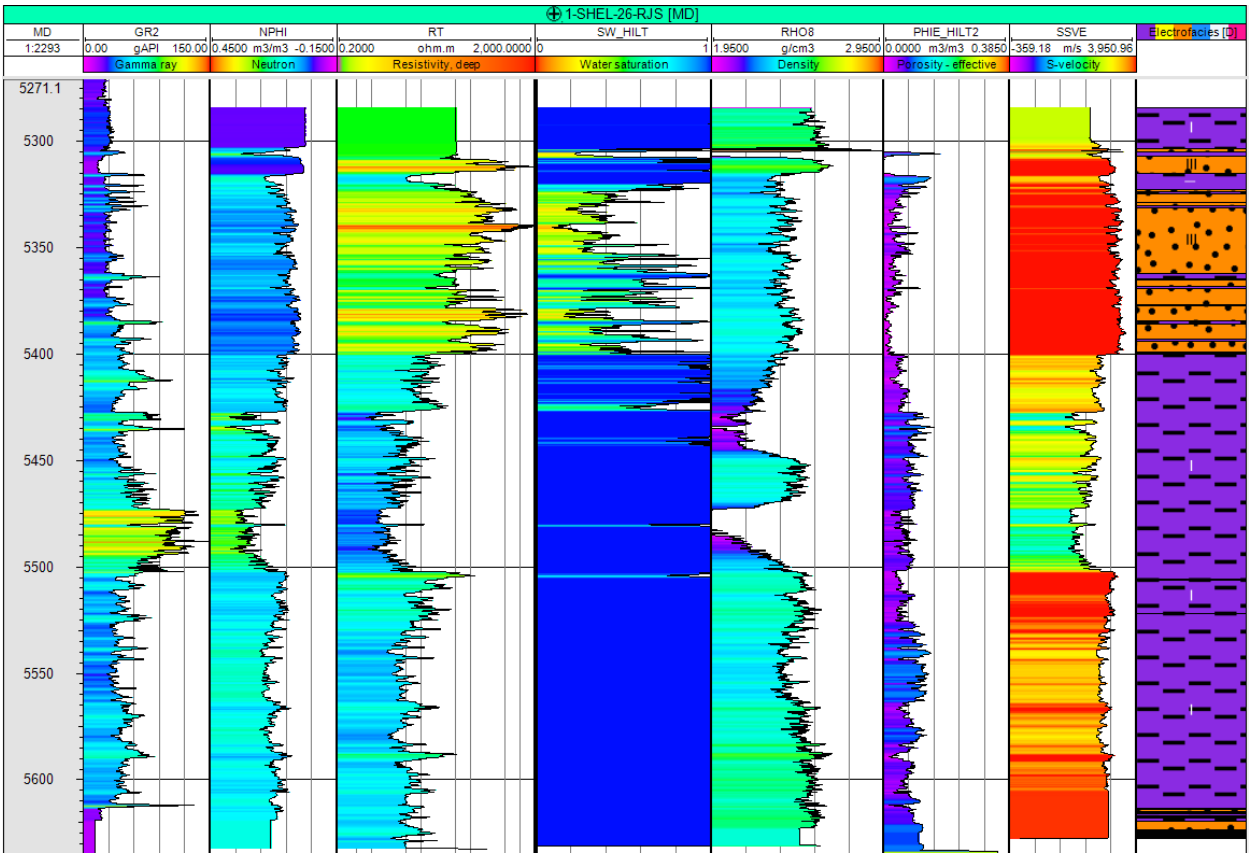



Figure 12. Well log section of the well log, 1-SHEL-26-RJS, with the electrofacies predited by the Model I.

Table 2. SVM - Confusion Matrix

| | | Predict | | | | |
|---|---|---|---|---|---|---|
| Real | | **123** | 0 | 0 | 2 | 0 |
| | | 0 | **25** | 0 | 0 | 0 |
| | | 0 | 1 | **113** | 7 | 4 |
| | | 0 | 0 | 2 | **85** | 0 |
| | | 1 | 0 | 8 | 3 | **32** |

Table 3. SVM - Classification Report

| EF | Precision | Recall | F1-score | Spp |
|---|---|---|---|---|
| 0 | 0.99 | 0.98 | 0.99 | 125 |
| 1 | 0.96 | 1.00 | 0.98 | 25 |
| 2 | 0.92 | 0.90 | 0.91 | 125 |
| 3 | 0.88 | 0.98 | 0.92 | 87 |
| 4 | 0.89 | 0.73 | 0.80 | 44 |
| **Acurracy = 0.93** | | | | |

Table 4. Neural Network - Confusion Matrix

| | | Predict | | | | |
|---|---|---|---|---|---|---|
| Real | | **143** | 0 | 0 | 1 | 0 |
| | | 0 | **24** | 0 | 0 | 0 |
| | | 0 | 0 | **119** | 1 | 0 |
| | | 1 | 0 | 5 | **68** | 0 |
| | | 2 | 0 | 3 | 0 | **39** |

Table 5. Neural Network - Classification Report

| EF | Precision | Recall | F1-score | Spp |
|---|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.99 | 144 |
| 1 | 1.00 | 1.00 | 1.00 | 24 |
| 2 | 0.94 | 0.97 | 0.96 | 120 |
| 3 | 0.96 | 0.93 | 0.95 | 74 |
| 4 | 0.97 | 0.89 | 0.93 | 44 |
| **Acurracy = 0.97** | | | | |

The Confusion matrix and Classification report generated in the Model II are shown in Tables 4 and 5:

It is possible to observe in Tables 3 and 5 an increase in the final value of all statistical parameters in the Model II when compared to the Model I. Final accuracy, F1 score, precision, and recall are higher for all electrofacies present in the work. These results revealed a greater robustness of Neural Networks and better matching with the data set of the research. When AI algorithms are working on a data set, it is important to check whether the algorithms are not in an overfitting or underfitting case. The overfitting case is verified when the algorithm presents a better performance

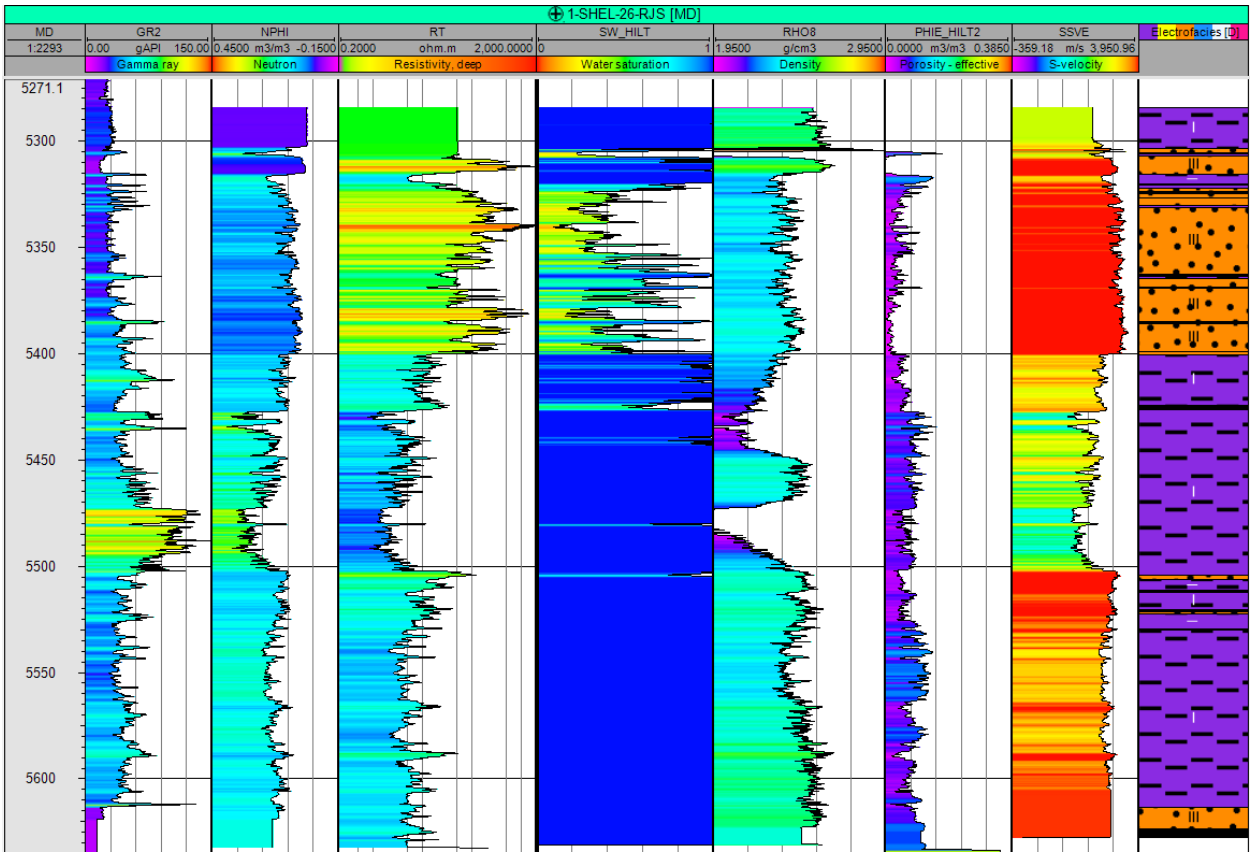



Figure 13. Well log section of the well log, 1-SHEL-26-RJS, with the electrofacies predited by the Model II.

in the training data set than test data set. In the case of overfitting, when the algorithm is submitted to the test data set, its performance is poor. The model I has presented a hit rate in the training data set of 93.7%, and a value of 92.9% in the test data set. The model II has presented hit rate in training and test data set of 96.9% and 96.6%. In both models, the underfitting and overfitting of models were verified and, in both cases, the hit rate found in the training data set was very similar with the hit rate in the test data set. This verification process indicates that the algorithms of the models are not in an overfitting case. The case of underfitting occurs when the algorithm performs poorly on the training and test data set. As has been shown with the statistical results, the algorithms applied in SVM model and Neural Network model are not in the case of underfitting.

In the current research, beyond the Confusion matrix method was utilized the k-Fold Cross-Validating to evaluate the accuracy. Using this method with five folds, k=5, were found an accuracy of 86.6% to the Model I, and an accuracy of 96.9% to the Model II. The results can be visualized in the Table 6.

According to the Tables 2, 3, 4, 5, and 6, it is possible to observe that the Model II (Neural Networks) presents a better performance at the Confusion matrix and K-Fold Cross-Validating when compared to the Model I (SVM).

Table 6. Models

| **Accuracy evaluation k-Fold Cross-Validating** | |
|---|---|
| Model I (SVM) | Model II (Neural Network) |
| 0.8307 | 0.9599 |
| 0.8492 | 0.9692 |
| 0.9661 | 0.9692 |
| 0.9507 | 0.9846 |
| 0.7314 | 0.9629 |
| **Mean ± Std** | **Mean ± Std** |
| 0.8656 ± 0.0959 | 0.9692 ± 0.0095 |

In relation to the Confusion matrix evaluation method, the Classification report (Table 3) of the Model I reveals a final accuracy about 93% and a worse performance in prediction of the electrofacies 0, 1, 2, and 4 when compared with the Model II. The only electrofacies that the Model I presents similar performance it is in electrofacies 0. The Confusion matrix of the Model II (Table 4) in general shows better performance. Individually, it is possible to note that the Model II does not present any error in the pre-

diction of electrofacies 01. A notable difference in performance of Models I and II can be visualized at electrofacies 04 (Tables 3, and 5), for the Model I the precision value is about 89%, while in the Model II the precision is about 97%, for recall the Model I has about 73% against about 89% of the Model II, in the F1-score the Model I presents performance about 80%, while the Model II has about 93%. Thus, according to the Confusion matrix statistical validation method, the better performance of the Model II (about 4% superior), about 97% against about 93%, can be explained mainly by observing the performance of both models in the prediction of the electrofacies 04.

From the products obtained by models I and II was created the index of heterogeneity (Eq. 12) of reservoirs zones. This parameter was created to measure the level of heterogeneity of the reservoir zone predicted by models I and II, Tables 9 and 10.

$$\mathbf{h} = \frac{TTN}{TTR + TTN} \tag{12}$$

Where, $TTN$ is the total thickness of non reservoir zone, $TTR$ is the total thickness of reservoir zone, and $h$ is the level of heterogeneity.

The Model I presents 13 sections of non reservoir zone inside the reservoir zone. The thirteen sections thickness of the Model I can be visualized in the Table 7. The minimum and maximum thickness of non reservoir zone inside of reservoir zone are 0.15 m, and 7.01 m respectively. The mean thickness of the non reservoir zone inside the reservoir zone to the Model I is 1.20 meters. The Model II presents 14 sections of non reservoir zone inside the reservoir zone. The fourteen sections thickness of the model II can be seen in the Table 8, it is possible to note the minimum and maximum thickness 0.15 m, and 4.42 m respectively. From the Table 8 it is possible to calculate the mean thickness of the non reservoir zone inside the reservoir zone to the Model II, 1.04 meters.

The tables 9, and 10 show the level of vertical heterogeneity in the well log section predicted by the models I, and II.

Table 7. Sections of non reservoir zones - Model I

| Depth Interval (m) | Section thickness (m) |
|---|---|
| 5304.74 - 5304.89 | 0.15 |
| 5305.20 - 5307.18 | 1.98 |
| 5315.41 - 5322.42 | 7.01 |
| 5323.64 - 5324.55 | 0.91 |
| 5328.51 - 5328.67 | 0.16 |
| 5330.34 - 5330.95 | 0.61 |
| 5362.35 - 5364.48 | 2.13 |
| 5364.79 - 5365.09 | 0.30 |
| 5368.44 - 5369.36 | 0.92 |
| 5376.98 - 5377.28 | 0.30 |
| 5384.60 - 5385.97 | 1.37 |
| 5392.98 - 5393.74 | 0.76 |
| 5398.92 - 5399.23 | 0.31 |
| Total | 16.90 |

Table 8. Sections of non reservoir zones - Model II

| Depth Interval (m) | Section thickness (m) |
|---|---|
| 5305.35 - 5306.87 | 1.52 |
| 5315.56 - 5319.98 | 4.42 |
| 5320.44 - 5322.27 | 1.83 |
| 5323.48 - 5324.55 | 1.07 |
| 5326.53 - 5326.84 | 0.31 |
| 5328.51 - 5328.67 | 0.16 |
| 5330.19 - 5330.95 | 0.76 |
| 5353.05 - 5353.32 | 0.27 |
| 5362.65 - 5364.48 | 1.83 |
| 5364.94 - 5365.09 | 0.15 |
| 5368.60 - 5369.20 | 0.60 |
| 5376.98 - 5377.13 | 0.15 |
| 5384.75 - 5385.97 | 1.22 |
| 5392.98 - 5393.28 | 0.30 |
| Total | 14.59 |

## 9 Discussion

According to the Confusion matrix Table 2 and Classification report 3, it is possible to notice that the SVM algorithm has the biggest statistical errors associated with the prediction of the electrofacies 04. The Confusion matrix, Table 2, shows that in the group of 44 samples, the SVM algorithm failed in 12 predictions. Among the 12 failed predictions, more than half (8) were classified as electrofacies 02. These results have a direct impact on the Classification report, causing the statistical scores of precision, recall, and f1-score to drop significantly in relation to the statistical scores of other electrofacies. In the research case, the input parameters values of the electrofacies 04 and 02 are very close, thus is comprehensive that the algorithm make mistakes. One way to try improving the statistics of the algorithm is to analyze the range of values of the parameters that make up the electrofacies. The accuracy level achieved with the detail level applied in electrofacies classification was better than the hoped to the research, therefore was not changed the range of the input parameters values for each electrofacies. But an alternative, in the case where the detail level can be decreased, is to merge the electrofacies that the algorithm is confusing in the prediction process, electrofacies 02 and 04. This process can increase the statistical scores, but on the other hand, the detail level of the electrofacies description decrease. With the current electrofacies configuration, the research achieved final values of hit rate above

Table 9. Heterogeneity level of reservoir zone in the Model I

| | |
|---|---|
| Total Thickness of reservoir zone | 96.62 m |
| Total Thickness of non-reservoir zone | 16.90 m |
| Nº of sections non reservoir zone | 13 |
| Vertical heterogeneity of the reservoir zone | 17.49% |

Table 10. Heterogeneity level of reservoir zone in the Model II

| | |
|---|---|
| Total Thickness of reservoir zone | 96.91 m |
| Total Thickness of non-reservoir zone | 14.59 m |
| Nº of sections non reservoir zone | 14 |
| Vertical heterogeneity of the reservoir zone | 15.06% |

90%. These results were considered acceptable and overcome the expectations, thus was not necessary to merge the electrofacies (02 and 04) and consequently to lose the current level of detail.

In the well section of the Model I, there is a reservoir zone predicted (electrofacies 02) in the depth interval of 5303.38 and 5400.00 meters, this interval has 96,62 meters of reservoir zone with non reservoir zones intercalated. But as can be seen in the figure 12, the interval of the reservoir zone is not homogeneous. Inside the reservoir zone there is 16,90 meters of non-reservoir zone divided in 13 sections. The thickness of 13 sections can be visualized in table 7. The index created to measure the level of heterogeneity of the reservoir zone (h = total thickness of non reservoir zone/total thickness of reservoir zone), found a value of 17.49% of heterogeneity in the reservoir zone 9. In the model II, the predicted reservoir zone (electrofacies 02) it is found in the depth interval of 5303.38 and 5400.29 meters, this interval has 96,91 meters of reservoir zone with non reservoir zones intercalated, figure 13. Inside the reservoir zone there is 14,49 meters of non-reservoir zone divided in 14 sections. The thickness of 14 sections can be visualized in table 8. The level of heterogeneity calculated to the reservoir zone was of 15.06%.

The knowledge about the heterogeneity level in a reservoir zone is an essential information because it allows to determine the better strategies in the well's operation moment.

In the Model II (Neural Networks), the Confusion matrix (Table 4) presents a matching between the predicted and real electrofacies considerably greater than the Confusion matrix of the Model I (Table 2). It is possible to verify by means of the Classification report, Table 5, that all statistical scores have increased in Model II. This occurs due to the better matching between the algorithm and the research data set, and also because the Neural Network is a more robust algorithm than SVM. In the Model II (Table 4), the problem of confusion between the prediction of electrofacies 02 and 04 can be ignored, because in the group of 44 samples labeled as electrofacies 04, only three were predicted as electrofacies 02.

The well logs sections of models I and II are very similar, among the five electrofacies available to fill the section, the models predicted the presence of only two, electrofacies 0 (purple) and 02 (orange). In the context of Gato do Mato oil field, these electrofacies presents two distinct zones, reservoir zone and non-reservoir zone. The electrofacies 0 has a greater content of clay that can be verified by the high values of gamma rays, it also has high values of density and water saturation representing a section with properties of non-reservoir zone. On the other hand, the electrofacies 02 have low values of gamma rays and water saturation, high values of resistivity, and moderate values of density, indicating properties of a reservoir zone.

In the well log section of the 1-SHEL-26-RJS, the most part is composed of non-reservoir zones (electrofacies 0), and a small portion is filled with reservoir zone (electrofacies 02). This behavior it is illustrated by models I and II in figures 12 and 13. It can be explained by the Gamma Rays values, it is possible to observe that the values of Gamma Rays are high in most part of the well. It is one of the most essential parameter used to distinguish between argilous zones and non-argilous zones. In research, this zones are responsible by auxiliary in the classification of reservoir and non-reservoir zones. Therefore, the main reason why well log 1-SHEL-26-RJS is mostly classified as a non-reservoir zone it is because of the values of Gamma rays. These high values lead the algorithms of models I and II to classify these portions as non-reservoir zones (electrofacies 0).

The main differences between the well log sections predicted by models I and II are that the well log section predicted by the Model I is more heterogeneous than well log section predicted by the Model II. In other words, the Model I has an electrofacies variation higher than Model II.

According to the results illustrated by SVM Model and Neural Network Model are clear that both models reaching a high-performance level. Both algorithms are capable to obtain a hit rate greater than 90%. Therefore both models can be used in electrofacies predictions with high level of reliability. When a performance comparison is made between the Models, the Neural Network Model shows to be about 4% better than SVM Model. In the well logs sections it is possible to observe that the classification made by the models is very similar, but the well log section pro-

duced by the Neural Network Model shows to be smoother and has smaller electrofacies variation than SVM Model. This behavior observed in the Neural Network Model is closer to the natural behavior verified in the oil field Gato do Mato, located in the Brazilian pre-salt.

## 10 Conclusions

The prediction of electrofacies in distinct data sets collaborates allowing us to know electrofacies present in different data sets in a short time period. How was demonstrated in the research, the models also are able to distinguish between the reservoir and non-reservoir zones. The results obtained with the models I and II can be considered promising once that the algorithms applied in models showed to be capable to predict with high precision the data set. In other words, the models I and II show a high potential to automate the process of electrofacies classification. This automation, if well calibrated as can be seen in models I and II can save time and resources for petroleum companies in large oil fields with a lot of well logs information. An advantage of models I and II to predict and classify the electrofacies is that the algorithms can quickly analyze several features, and thus, to predict patterns in regions difficult to be find by human eyes. Another advantage is that the input features, once associated with the corresponding electrofacies, can be used to generalize the overall reservoir. Therefore, it becomes possible the data set that represents a few inches of the surrounding well can be interpolated by the wells in the oil field and, consequently, to generate a model of the entire reservoir.

The models I and II showed a convergence in their predictions. This is a good sign because reinforcement the predictions of the models. Beyond that was possible to associate reservoir and non reservoir zones to the electrofacies predicted. Due to the characteristics of the Geophysical well logs response was possible to associate the electrofacies 02 with a reservoir zone, and the electrofacies 0 with a non reservoir zone. The models I and II has similar level of vertical heterogeneity. The difference about 2% between the vertical heterogeneity in models reinforce the convergence of the predicted electrofacies by models I and II. The Model II has reveal high values of final accuracy in both methods of evaluation, Confusion matrix and k-Fold Cross-Validation. In addition, the Model II showed to be more consistent in the five predictions made by the k-Fold Cross-Validation. But the higher robustness of the Model II in relation to the Model I has a price. While the Model I has an execution time of 0.096s, the Model II presents an execution time of 41.45s. In other words, the Model I is about 432 times more quickly than Model II. Therefore, whether the difference of about 4% (97% Model II - 93% Model I) it is not a problem, then Model I is more recommended, once time that it is as good as Model II, and about 432 times more quickly.

The current research approached themes of high importance in the modern scenario, the union between well logs data set and machine learning algorithm have shown a success combination. Therefore, the work contribute for the literature of geosciences and AI algorithms as a practical case of AI algorithms automating an important process in the petroleum geology.

The Models presented in the research open doors to other applications in the O&G sector. The own models applied in the current research can be recalibrated for predictions of others sectors. Once the point is reached the models presented in the research can be recalibrated to reach similar or better performance. There are researches being developed by the research group with the main goal to explore other possibilities to apply the models developed in the current research in other areas of O&G. The future of O&G and other technological sectors has shown that the combination of AI algorithms and human knowledge is a long-term marriage.

## 11 Acknowledgments

I wish to thank Alessandro Batezelli for help in my master and in current work. I wish to thank the Brazilian National Agency of Petroleum for providing the data set that was used in this work. I wish to thank the Faculty of Mechanical Engineering(FEM) and Petroleum Engineering Department (DEP) to contribute with the all physical structure, buildings and laboratories, that enabled the project to be carried out.

## References

[1] Tristan Euzen, Eric Delamaide, Tom Feuchtwanger, Kim D Kingsmith, et al. Well log cluster analysis: an innovative tool for unconventional exploration. In *Canadian unconventional resources and international petroleum conference*. Society of Petroleum Engineers, 2010.

[2] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[3] Donald F Specht et al. A general regression neural network. *IEEE transactions on neural networks*, 2 (6):568–576, 1991.

[4] Stephen I Gallant and Stephen I Gallant. *Neural network learning and expert systems*. MIT press, 1993.

[5] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[6] Andre Torres and Alessandro Batezelli. Applying supervised machine learning model to classify electrofacies in a brazilian pre-salt wellbore. 2020. doi:https://doi.org/10.48072/2525-7579.rog.2020.014.

[7] Weiming Xiang, Hoang-Dung Tran, and Taylor T Johnson. Reachable set computation and safety verification for neural networks with relu activations. *arXiv preprint arXiv:1712.08163*, 2017.

[8] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[9] O Serra and H Abbott. The contribution of logging data to sedimentology and stratigraphy.(spe paper 9270.) paper presented at the 55th annual fall technical conference and exhibition of the society of petroleum engineers of aime. *Dallas, TX*, pages 21–24, 1980.

[10] Bishnu Kumar and Mahendra Kishore. Electrofacies classification–a critical approach. In *6th International Conference & Exposition on Petroleum Geophysics, New Delhi, India*, pages 822–825, 2006.

[11] Gary Nichols. *Sedimentology and stratigraphy*. John Wiley & Sons, 2009.

[12] Andrew D Miall. *The geology of stratigraphic sequences*. Springer Science & Business Media, 2010.

[13] Bishnu Kumar and Mahendra Kishore. Electrofacies classification—a critical approach. In *6th international conference and exposition on petroleum geophysics, Kolkata, India*, pages 822–825, 2006.

[14] Hung Kiang Chang, Mario Luis Assine, Fernando Santos Corrêa, Julio Setsuo Tinen, Alexandre Campane Vidal, and Luzia Koike. Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na bacia de santos. *Revista Brasileira de Geociências*, 38(2 suppl):29–46, 2008.

[15] Paulo Roberto Schroeder Johann, Rubens Caldeira Monteiro, et al. Geophysical reservoir characterization and monitoring at brazilian pre-salt oil fields. In *Offshore Technology Conference*. Offshore Technology Conference, 2016.

[16] Ildo L Sauer. O pré-sal e a geopolítica e hegemonia do petróleo face às mudanças climáticas e à transição energética. *Recursos Minerais do Brasil*, 2016.

[17] Mitsuru Arai. Aptian/albian (early cretaceous) paleogeography of the south atlantic: a paleontological perspective. *Brazilian Journal of Geology*, 44(2): 339–350, 2014.

[18] Jobel Lourenço Pinheiro Moreira, Cláudio Valdetaro Madeira, João Alexandre Gil, Marco Antonio Pinheiro Machado, et al. bacia de santos. *Boletim de Geociencias da PETROBRAS*, 15(2):531–549, 2007.

[19] Mario Carminatti, Breno Wolff, Luiz Gamboa, et al. New exploratory frontiers in brazil. In *19th World Petroleum Congress*. World Petroleum Congress, 2008.

[20] Hamed Kiaei, Yousef Sharghi, Ali Kadkhodaie Ilkhchi, and Mehrangiz Naderi. 3d modeling of reservoir electrofacies using integration clustering and geostatistic method in central field of persian gulf. *Journal of Petroleum Science and Engineering*, 135:152–160, 2015.

[21] Robert L Folk. Practical petrographic classification of limestones. *AAPG bulletin*, 43(1):1–38, 1959.

[22] RL Folk. Spectral subdivision of limestone types: American association of petroleum geologists memoir, v. 1. 1962.

[23] Robert J Dunham. Classification of carbonate rocks according to depositional textures. 1962.

[24] Vladimir N Vapnik. The nature of statistical learning. *Theory*, 1995.

[25] Vladimir Vapnik. The support vector method of function estimation. In *Nonlinear modeling*, pages 55–85. Springer, 1998.

[26] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

[27] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[28] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[29] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152.

[30] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[31] Shun-ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[32] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[33] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62, 1997.

[34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[35] Lei Ba. *Adaptive dropout for training deep neural networks*. PhD thesis, 2013.

[36] Guifang Lin and Wei Shen. Research on convolutional neural network based on improved relu piecewise activation function. *Procedia computer science*, 131:977–984, 2018.

[37] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[38] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

[39] Digvijay Boob, Santanu S Dey, and Guanghui Lan. Complexity of training relu neural network. *Discrete Optimization*, page 100620, 2020.

[40] Meiqi Wang, Siyuan Lu, Danyang Zhu, Jun Lin, and Zhongfeng Wang. A high-speed and low-complexity architecture for softmax function in deep learning. In *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 223–226, 2018. doi:10.1109/APCCAS.2018.8605654.

[41] Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. Efficient softmax approximation for gpus. In *International Conference on Machine Learning*, pages 1302–1310. PMLR, 2017.

[42] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[44] Nitish Srivastava. Improving neural networks with dropout. *University of Toronto*, 182(566):7, 2013.

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

[46] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[47] Jeff Miller. Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, 43(4):907–912, 1991.

[48] David C Howell, M Rogier, V Yzerbyt, and Y Bestgen. Statistical methods in human sciences. *New York: Wadsworth*, 721, 1998.

[49] Stephen Kokoska and Daniel Zwillinger. *CRC standard probability and statistics tables and formulae*. Crc Press, 2000.

[50] Denis Cousineau and Sylvain Chartier. Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67, 2010.

[51] Robert Dawson. How significant is a boxplot outlier? *Journal of Statistics Education*, 19(2), 2011.

[52] Jorma Laurikkala, Martti Juhola, Erna Kentala, N Lavrac, S Miksch, and B Kavsek. Informal identification of outliers in medical data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology*, volume 1, pages 20–24. Citeseer, 2000.

[53] HP Vinutha, B Poornima, and BM Sagar. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences*, pages 511–518. Springer, 2018.

[54] AH Joarder and M Firozzaman. Quartiles for discrete data. *Teaching Statistics*, 23(3):86–89, 2001.

[55] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[56] Irad Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer, 2005.

[57] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9):1635–1647, 2004.

[58] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of rnn for outlier detection in data mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 709–712. IEEE, 2002.

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[60] Hatem A Fayed and Amir F Atiya. Speed up grid-search for parameter selection of support vector machines. *Applied Soft Computing*, 80:202–210, 2019.

[61] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, 14(4):1502, 2016.

[62] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[63] Yuhong Yang et al. Consistency of cross validation for comparing regression procedures. *Annals of Statistics*, 35(6):2450–2473, 2007.

[64] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. springer series in statistics. In ∴. Springer, 2001.

[65] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Statistical learning. In *An Introduction to Statistical Learning*, pages 15–57. Springer, 2013.

[66] Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[67] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

[68] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella. The'k'in k-fold cross validation. In *ESANN*, pages 441–446, 2012.

[69] Carlos EM dos Anjos, Manuel RV Avila, Adna GP Vasconcelos, Aurea M Pereira Neta, Lizianne C Medeiros, Alexandre G Evsukoff, Rodrigo Surmas, and Luiz Landau. Deep learning for lithological classification of carbonate rock micro-ct images. *Computational Geosciences*, pages 1–13, 2021.

[70] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011.